Panopticon: Advancing Any-Sensor Foundation Models for Earth Observation

Supplementary Material

A. Why "Panopticon"?

The idea of the *panopticon* was first suggested by the philosopher Jeremy Bentham [2] as an ideal model for efficient prison design, where a single person could watch over an entire prison. Michel Foucault later reinterpreted it as a powerful metaphor for repressive systems of power, control and surveillance in modern societies [9], something that is perhaps even more relevant today with the proliferation of digital surveillance technologies.

We are well aware of the term's loaded history and controversial connotations; so why choose it? We want to co-opt this term and flip its meaning, using it as a metaphor for systems that can keep a watchful and benevolent eye over our planet. Instead of surveilling people, Panopticon(s) can observe Earth itself—its changing landscapes, ecosystems, and climate patterns.

The beauty of our model is that, like the original panopticon concept, it provides comprehensive visibility from a single vantage point. But unlike Bentham's prison design, our goal is not control and fear, but rather understanding and monitoring. There is also a technical parallel that we find fitting: the original panopticon was designed to see everything from a central position, regardless of where attention was directed. Similarly, our model can "look" through any sensor configuration without needing specific adaptations—a sensor-agnostic vision that mirrors the all-seeing nature of the conceptual panopticon, but repurposed for planetary good.

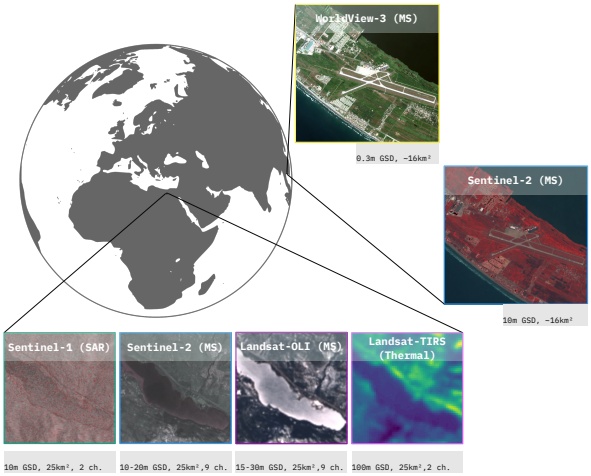
B. Code and data availability

All data loaders and model code will be contributed to the TorchGeo library [18] for reproducibility and ease of future experimentation.

C. Datasets

C.1. Pre-training datasets

We organize the four pretraining datasets as shown in Table 1 of the main text by geographical footprint, where a footprint is defined by images having an exact geo-reference match. During pretraining, from each footprint, we randomly sample a number of "snapshots" as shown in Fig. 1.

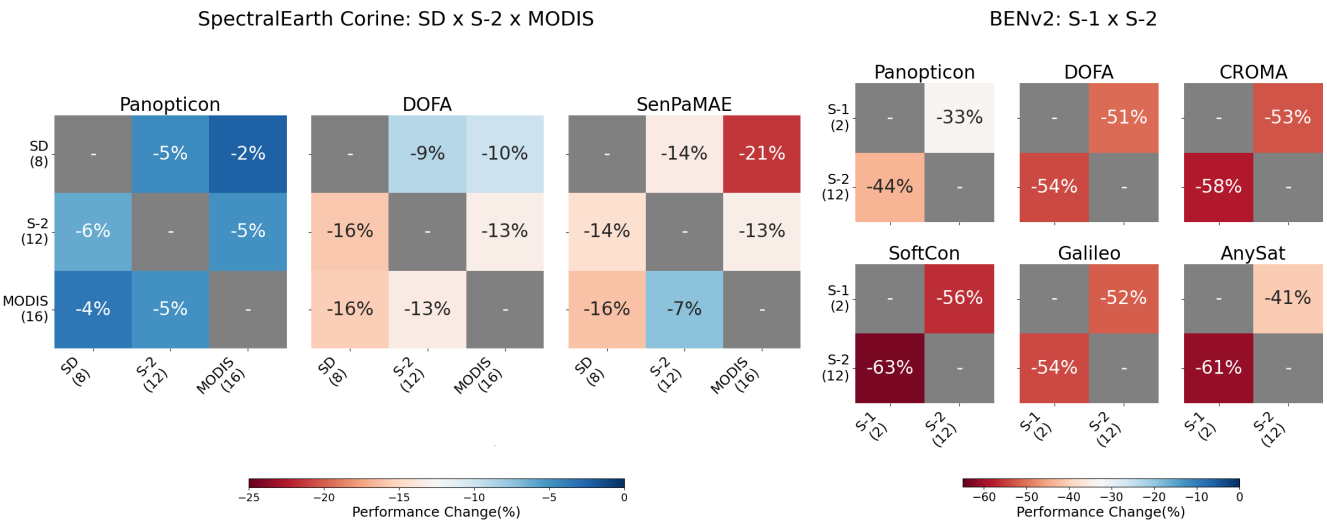


Supp. Fig. 1. Snapshots: examples of snapshots taken from distinct footprints. Our pretraining dataset consists of various sensor modalities, channels, GSDs, scales, and timesteps acquired from across the Earth's surface. Different channels, GSDs, and footprint sizes provide information about different attributes of the geospatial objects. Note: some images are shown in false colors to enable mapping from non-visible spectra.

fMoW fMoW [5] consists of data from 4 MS satellite sensors, QuickBird-1 and GeoEye-2 having 4 channels (RGB, NIR), while WorldView-2 and -3 have 8 channels, with GSDs typically ranging from 1–2 m. Additionally, the dataset also includes pansharpened versions of these same images in RGB which have a GSD < 1 m. This dataset was chosen for its global spatial coverage, wide spectral coverage[1] and very low GSD values, along with extensive functional coverage of human modified land cover types. Moreover, this dataset also represents a large variance in time of capture, off-nadir angles, both of which affects illumination of targets. Since spatial footprints were available for all images, we generate an geographically indexed version of this dataset, which will be released with the rest of the code. Additionally, we remove images greater than 1024 px, which are typically the pansharpened RGB images, to reduce memory overheads.

fMoW-Sentinel fMoW-Sentinel [7] was created to be an exact copy of the locations captured by fMoW, but with Sentinel-2 imagery. We created a combined dataset from fMoW and fMoW-Sentinel by indexing by footprint and sensor type. Together, these two datasets capture surface properties from five separate sensor platforms between 2002 and 2022, providing a lot of natural variation for a given footprint. Finally, the footprints of each image vary

[1] including "non-standard" bands, i.e. those with spectral coverage outside those of the popular open-data sources such as Landsat and Sentinel series.



Supp. Fig. 2. Cross-sensor invariance. In addition to train/val/test splits, we also split datasets across sensors to explicitly test sensor invariance. y-axes on the heatmap represent training sensors and x-axes, test sensors. The diagonals represent same-sensor for train and test, and are grayed-out, while off-diagonal elements represent cross-sensor results with values expressed as percentage differences from the diagonals. This enables visualization of the change in cross-sensor performance relative to same sensor, expressed in (negative) percentages. **Left**: splits are across synthesized sensors from the HS EnMAP sensor using spectral convolution - MODIS, Sentinel-2 and Planet SuperDove, for any-sensor models. **Right**: reBEN; splits are implemented across Sentinel-1 and Sentinel-2 sensors, for all any-sensor and many-sensor models, except SenPaMAE which cannot process SAR imagery. Panopticon consistently outperforms other any-sensor models. Note that the value ranges differ for the two sub-figures.

tremendously, from 0.2 to 25 km^2, providing a large range of features at different scales. This combined dataset consists of 89,666 unique footprints, where each footprint can have dozens of images across these sensors.

MMEarth Multi-modal Earth (MMEarth) [14] was released as a paired dataset of multiple modalities that include Sentinel-1 SAR, Sentinel-2 MS, elevation, and other paired modalities. Most importantly, the Sentinel-2 data included multiple processing levels (L1C and L2A), and Sentinel-1 data was captured in all 4 polarization combinations (VV, VH, HH, HV) and in both orbits (ascending, descending). This data was primarily included to model the effects of polarization and orbit for SAR data and processing levels for optical. Moreover, this extensive dataset comprises of 1,239,937 unique footprints equitably distributed according to land cover types, each of which providing a pair of S1 and S2 images.

SpectralEarth SpectralEarth [3] is the largest open source hyperspectral dataset available at the time of writing comprising of 450 K patches sampled globally by the EnMAP satellite [11], made available by the German Aerospace Center (DLR). This dataset additionally provides four downstream benchmark tasks using data from the same sensor, but utilizing non-overlapping patches, separate from pretraining. This dataset was included for its rich spectral diversity, enabling the model to learn HS characteristics and simulate any arbitrary multispectral band. SpectralEarth provides 415 K unique footprints, each of which provides a single HS image of 202 bands.

SatlasPretrain SatlasPretrain consists of 30 TB of imagery across Sentinel-1, Sentinel-2, Landsat-9, and NAIP sensors. We utilize only the first three, as NAIP geographic coverage is limited to the United States and spectrally consists of only RGB and NIR bands. We created a unified indexed dataset comprising of 768,800 unique footprints, where each footprint can contain up to 3 sensor images taken across 2022. It is also the only large pretraining dataset to contain thermal images from the Landsat 9-TIRS sensor.

C.2. Evaluation datasets

The utilize the following benchmark task and datasets. Where possible, we utilized existing Python libraries such as TorchGeo [18] and GEO-Bench [13]. For a complete list of datasets, please consult Tab. 1. Most datasets are implemented using the TorchGeo library [18], when available.

In the following, we outline any modifications we make to standard datasets:

SpaceNet We utilize the SpaceNet 1 dataset from TorchGeo, which is a building footprint segmentation task over the city of Rio de Janeiro with 8 band MS images and 3 band pansharpened RGB images captured by WorldView 2. We utilize only the 8-band images. Since the original dataset is only available with labels for the training set, we randomly split the dataset into training, validation and test splits with a 80:10:10 ratio.

D. Additional Results

D.1. Sensor invariance

We explicitly test for a model's ability to generate stable representations regardless of the sensor input. To do this, we implement an additional split dimension on datasets that have paired sensors, splitting on the sensor. We use the 10 EnMAP-Corine [3] dataset where we employ spectral convolution to simulate MODIS, Sentinel-2, and Planet SuperDove sensors (Fig. 2 (left)) and the 12 reBEN [6] dataset that has paired imagery from Sentinel-1 and Sentinel-2 (Supp. Fig. 2 (right)). Models are trained on the training split of the dataset using a single sensor, while being validated and tested on a corresponding split of the dataset, using data from a different sensor. We then plot the relative difference to the same-sensor setting as shown on the off-diagonal cells. This allows us to validate how close representations generated from one sensor are to ones generated from a different sensor on the same dataset and prediction objective. Ideally, the off-diagonal cell values are close to 0%, i.e. similar to the same-sensor setting. This is the case for Panopticon for the Corine dataset in Fig. 2 (left), where other any-sensor models struggle to generalize, especially when training on fewer bands (SD with 8 bands) and testing on sensors with more bands (MODIS with 16 bands). This effect is less pronounced in Fig. 2 (right), where the domain shift is very strong going from optical to SAR and vice-versa. However, even in this setting Panopticon outperforms all any-sensor and many-sensor models by 18% and 14%, for the S1→S2 and S2→S1 settings on average, respectively.

Absolute values for these experiments along with additional results on the fMoW dataset split according to three included sensors, can be seen in Supp. Fig. 3.

D.2. Geo-Bench

We present the full results on Geo-Bench in Table 4 and Table 5.

E. Utilizing complete spectral information

In the field of any-sensor models, DOFA [21] uses a hypernetwork and the mean of the SRF to learn spectral encodings per channel, while SenPaMAE [16] utilizes the full SRF. We experimented with the following mechanisms to incorporate SRF and/or bandwidth information per channel, in addition to the mean wavelength.

Spectral integrated positional encoding To achieve sensor agnosticism, we introduce a novel spectral integrated positional encoding (sIPE) that leverages known sensor characteristics. Building on the channel-wise attention mechanism from Nguyen et al. [15], we model each sensor's SRF as an un-normalized Gaussian kernel characterized by its mean, μ and standard deviation, σ. This was inspired by noticing that such a Gaussian kernel provides a relatively good fit for SRFs, such as Sentinel-2A and Landsat-9 OLI sensors as shown in Fig. 4. We also experimented with Epanechkinov kernel fitting, which provided better R^2 results, but since Gaussians are well understood and have closed-form solutions, they tend to be easier to work with. Hence, we model the SRF of a channel with the parameters $\{\mu, \sigma\}$ of a Gaussian fit. We then extend absolute positional embeddings [19] to incorporate this spectral information through integration against sinusoidal basis functions

$$\begin{aligned} \text{PE}_{\text{spectral}}(\mu, \sigma, 2i) &= \int e^{-\frac{(\mu-\lambda)^2}{2\sigma^2}} \cdot \sin(\omega_i \lambda) \cdot \mathrm{d}\lambda, \\ \text{PE}_{\text{spectral}}(\mu, \sigma, 2i+1) &= \int e^{-\frac{(\mu-\lambda)^2}{2\sigma^2}} \cdot \cos(\omega_i \lambda) \cdot \mathrm{d}\lambda, \end{aligned} \tag{1}$$

where $\omega = \frac{1}{10000^{\frac{2i}{D}}}$ and $i \in (0, D]$.

We derive the closed-form solution for Eq. (1) as follows:

$$\begin{aligned} \text{PE}_{\text{spectral}}(\mu, \sigma, 2i) &= \sigma\sqrt{2\pi} \cdot e^{-\frac{\omega_i^2 \sigma^2}{2}} \cdot \sin(\omega_i \mu), \\ \text{PE}_{\text{spectral}}(\mu, \sigma, 2i) &= \sigma\sqrt{2\pi} \cdot e^{-\frac{\omega_i^2 \sigma^2}{2}} \cdot \cos(\omega_i \mu). \end{aligned} \tag{2}$$

Eq. (2) allows us to efficiently generate spectral PEs that are added to channel tokens. Our hypothesis was that while the query learns how best to fuse spectral tokens across channels, it can only do so based on the extracted low-level features within those patches. Adding the sIPE to the patch tokens provides a spectral inductive bias to each token relative to its central wavelength and bandwidth, effectively grounding the patch token to its physical capture attributes.

Spectral convolution Inspired by King et al. [12], we employ spectral convolution [4] as a spectral augmentation for HS inputs. Given a source HS image comprised of i channels, $x^s(\lambda)$, and its spectral response function SRF_s, a spectral convolution R is defined as the integral of the product of x^s and SRF_s, normalized by the integral of its SRF. To generate arbitrary MS channels with unknown SRFs, we model their SRF, SRF_t as un-normalized Gaussian kernel with a mean wavelength, λ_t, and standard deviation, σ_t. Through this process, we are able to generate novel multi-spectral

Index	Name	Name used in this paper	Modifications
1	Corine [SuperDove]		Spectral convolution from EnMap to Planet Superdove
2	Corine [MODIS]		Spectral convolution from EnMap to MODIS
3	Hyperview [SuperDove]		Spectral convolution from Intuition to SuperDove
4	Hyperview [MODIS]		Spectral convolution from Intuition to MODIS
5	TropicalCyclone	TC	TorchGeo, we use 10% of train
6	DigitalTyphoon	DT	TorchGeo, we use 10% of train
7	SpaceNet 1	SpaceNet 1	Randomly split the train set into train, val, and test (80:10:10)
8	EuroSAT	m-eurosat	GEO-Bench
9	BrickKiln	m-brick-kiln	GEO-Bench
10	EnMAP-Corine	Corine	Original 202 band dataset
11	RESISC45	RESISC	GEO-Bench
12	reBEN-S2		
13	reBEN-S1		
14	ForestNet	m-forestnet	GEO-Bench
15	So2Sat	m-so2sat	GEO-Bench
16	PV4Ger (cls.)	m-pv4ger	GEO-Bench
17	PV4Ger (segm.)	m-pv4ger-seg	GEO-Bench
18	Cheasapeake Landcover	m-chesapeake-landcover	GEO-Bench
19	Cashew Plantation	m-cashew-plantation	GEO-Bench
20	SA Crop Type	m-SA-crop-type	GEO-Bench
21	NZ Cattle	m-nz-cattle	GEO-Bench
22	NEON Tree	m-NeonTree	GEO-Bench
23	fMoW		Subset to WV2+WV3 sensors

Table 1. Summary of all evaluation datasets and the modifications made to them.

views from any HS source, extensively expanding the spectral augmentation capabilities of this framework. We hypothesize that this combination may prove to add useful spectral inductive biases to the model.

Evaluation To evaluate these design choices, we run the evaluation suite on a model trained according to the specifications laid out in Sec 3.5, i.e. identical to the model described in the main paper. We call this model Panopticon-IPE. The results are shown in Tab. 2.

Comparing these results to Tab 2 & 3, we see that Panopticon-IPE is relatively close in performance to Panopticon, and in some cases (TropicalCyclone) even excels. However, on average this model performs worse than our default settings, and as a result, we did not include these methods in the main paper. We leave these findings for the benefit of future researchers.

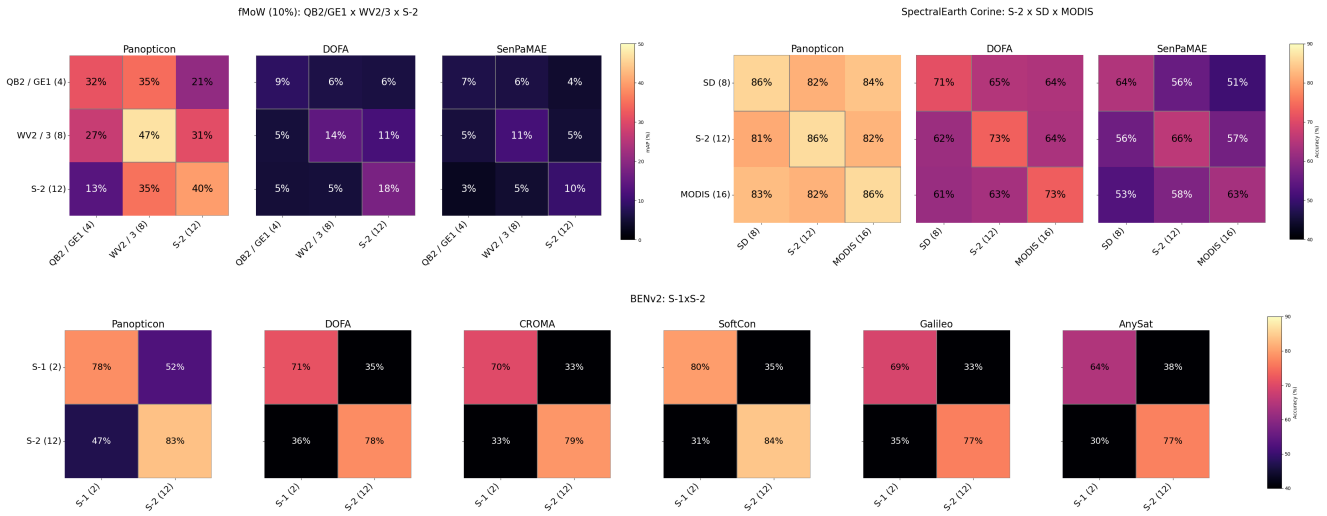
F. Technical details

F.1. Pre-training

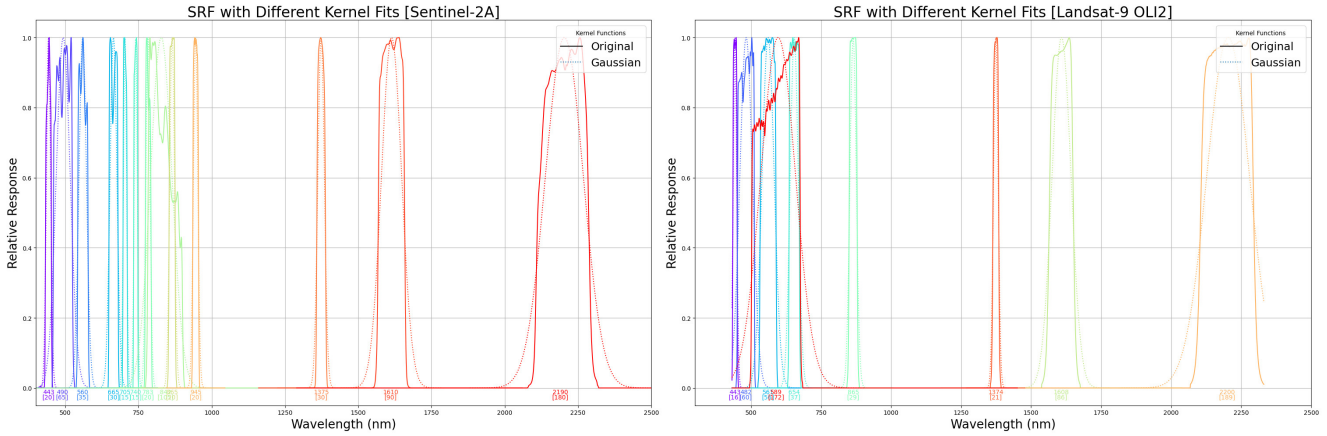
Implementation of Channel Attention We implement the initial shared 2D convolution of the PE with a $1 \times p \times p$ 3D convolution [1], where p is the patch size. Note that images within a batch can originate from different sensors in our pipeline and, thus, the number of channels is not consistent within a batch. To efficiently compute the cross attention for batched inputs, we hence employ padding and masking.

Hyperparameters We follow most of the DINOv2 configuration with the following changes: We multiply the learning rate of the ViT blocks in the backbone by 0.2, reduce the iBOT loss weight to 0.1, and remove the KoLeo regularizer. We performed early off-the-record ablations on these choices. Apart from that, and for both stages 1 and 2, we train for 70 artificial epochs with 1250 iterations each, with the AdamW optimizer, a base learning rate of 5e-4, standard learning rate scaling $\mathrm{lr} \cdot \mathrm{bsz}/256$, and a linear learning rate warmup for 5 epochs followed by a cosine decay to a minimum 1e-6 learning rate.

Metrics For details on the average metrics defined in the ablations, see Tab. 3.



Supp. Fig. 3. Cross-sensor invariance (absolute values). In addition to train/val/test splits, we also split datasets across sensors to explicitly test sensor invariance. y-axes on the heatmap represent training sensors and x-axes, test sensors. The diagonals represent same-sensor for train and test, while off-diagonal elements represent cross-sensor results with all values expressed as absolute performance values as percentages. **Left**: splits are across QuickBird-2 (QB2) / GeoEye-1 (GE1) which are RGB-NIR. **Right**: splits are across synthesized sensors from the HS EnMAP sensor using spectral convolution - MODIS, Sentinel-2 and Planet SuperDove, for any-sensor models. **Bottom**: reBEN; splits are implemented across Sentinel-1 and Sentinel-2 senors, for all any-sensor and many-sensor models, except SenPaMAE which cannot process SAR imagery. Panopticon consistently outperforms other models in both settings. Note that the value ranges differ for each sub-figure.



Supp. Fig. 4. Spectral response function (SRF) fitting for Sentinel-2 (left) and Landsat 9 (right).

F.2. Evaluation

All tasks were executed on either a 40 GB A100 or a 48 GB L40S GPU. The batch size is optimized for each task and model to maximize GPU memory and the base learning rate lr is scaled by the batch size according to the linear scaling rule $\text{lr} \cdot \text{bsz}/256$ [10]. In our evaluations, we use the following tasks types.

kNN k-nearest neighbors (kNN) with $k = 20$ and temperature 0.07 following Reed et al. [17].

Linear probing We mainly follow the implementation of DINOv2 and sweep multiple extraction methods. In particular, we extract the tokens from the last one or four transformer blocks and aggregate them by concatenating the cls tokens, average pooling, or the default aggregation suggested by the specific model at hand. For each aggregation, we sweep the 13 base learning rates 1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.1, 0.5, 1, 5, and 10, resulting in $2 \cdot 3 \cdot 13 = 78$ runs. Note that we use an extension of the DINOv2 implementation that computes all these different configurations simultaneously from a single forward pass

Table 2. Performance metrics for Panopticon-IPE across various datasets

Dataset	Score
Classification (Accuracy %)	
m-eurosat	96.1
m-brick-kiln	95.8
m-pv4ger-cls	95.8
RESISC45	91.2
m-forestnet	54.0
Segmentation (mIoU %)	
m-pv4ger	95.4
m-nzcattle	92.8
spacenet1	90.3
m-neontree	79.6
m-chesapeake	78.0
m-cashew	59.1
m-sacrop	52.3
Multi-Label Classification (mAP %)	
Corine-MODIS	80.0
Corine-SuperDove	79.8
Regression (MSE)	
DigitalTyphoon	0.93
Hyperview-MODIS	0.35
Hyperview-SuperDove	0.35
TropicalCyclone	0.28

of the backbone. We train for 50 epochs with a 0.9 momentum Stochastic Gradient Descent optimizer. We also add a trainable batch normalization before the linear head. For the cross-sensor evaluations in Fig. 2 and Fig. 3, we only extract tokens from the last transformer block to ease compatibility issues across models that use different encoders for SAR and optical modalities.

Linear probing with re-initialized patch embedding We replace the patch embedding of the model with a randomly initialized 2D convolution layer with the correct number of channels of the dataset at hand. We add a trainable batch norm and linear head to the encoder, unfreeze the new patch embedding and freeze the remaining encoder. We fix the feature aggregation to the default one suggested by the model authors and sweep the base learning rates 0.01, 0.001, 0.0005, and 0.0001 with 50 epochs each, stochastic gradient descent optimizer and 5 epochs of learning rate warmup.

AnySat and Galileo presented unique challenges due to the way they implement modality specific encoders. For AnySat, we pick a specific modality, and retrain the 2D convolution layer of that branch; we picked the NAIP encoder since it does not implement temporal attention, which consumes significant memory and compute. For Galileo, this was not possible since they use a channel grouping where each group produces a set of independent tokens, unlike other models where the channel dimension is collapsed. To replace this, we would have to create a new group with a custom number of channels which would break how Galileo processes tokens in groups. Furthermore, Galileo employs FlexiPatchEmbed, which is trained by randomizing the patch size during pretraining. To properly train this module, we would need to mimic that during evaluation, which was beyond the scope of the evaluation phase of this paper. Therefore, we omitted evaluating Galileo in tests that required retraining the patch embed module for domain adaptation.

Semantic segmentation We freeze the backbone and add trainable standard modules from the MMSegmentation library [8]. In particular, we use a Feature2Pyramid as neck, a UPerNet decoder and an auxiliary FCNHead. We sweep the base learning rates 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6 and utilize the AdamW optimizer for 50 epochs with no learning rate warmup.

Model adaptations During evaluation, we picked the image size that the model was natively trained on to maximize that model's ability to produce representations.

G. Additional ablations

Spectral augmentation In line with the DINO ablation on scales in random resized crops, we ablate our spectral size as $(1, s), (s, 13)$ for $s \in \{2, 4, 8, 13\}$. In Table 6, we see that $s = 4$ performs best.

Agg.	Id	Dataset	Task details	Metric
MS_{acc}	1	m-eurosat	kNN, $k = 20, T = 0.07$	top-1 micro accuracy
	2	m-eurosat without RGB channels	kNN, $k = 20, T = 0.07$	top-1 micro accuracy
SAR_{acc}	3	m-eurosat-SAR [20]	kNN, $k = 20, T = 0.07$	top-1 micro accuracy
	4	m-eurosat-SAR	linear probing for 10 epochs	top-1 micro accuracy
Sim_{mAP}	5	Corine-4	linear probing on 10% subset for 10 epochs	top-1 micro multilabel average precision
	6	Corine-12	linear probing on 10% subset for 10 epochs	top-1 micro multilabel average precision
Avg	1–6			
	7	RESISC45	kNN, $k = 20, T = 0.07$	top-1 micro accuracy
	8	m-eurosat only RGB channels	kNN, $k = 20, T = 0.07$	top-1 micro accuracy

Table 3. Metrics used to inform design decisions and reported in the ablation section of the main text. The aggregation metric is computed as the average of all its metrics. Corine-n denotes the Corine dataset subsampled to n fixed randomly-selected channels.

	m-brick-kiln (S2)	m-eurosat (S2)	m-forestnet (L8)	m-pv4ger (RGB)	m-so2sat (S2)	reBEN (S2)
DINOv2	**97.5**	95.5	53.5	**97.5**	60.8	80.1
CROMA	94.5	91.1	-	-	53.5	79.4
SoftCon	94.9	92.2	-	-	52.1	80.6
AnySat	90.3	87.6	50.9	92.8	42.5	76.8
Galileo	93.1	88.6	-	-	54.2	76.5
SenPaMAE	83.9	77.5	33.5	87.1	33.7	63.8
DOFA	95.8	92.9	53.2	97.4	54.2	78.8
Panopticon	96.7	96.4	**56.3**	96.4	61.7	**83.9**

Table 4. Linear probing on GEO-Bench classification datasets and reBEN. We report micro accuracy for single-label and mean average precision for multi-label datasets in percentages.

	m-cashew (S2)	m-chesapeake (RGBN)	m-neontree (RGB)	m-nzcattle (RGB)	m-pv4ger (RGB)	m-sacrop (S2)
DINOv2	**65.9**	<u>78.5</u>	**80.9**	92.7	**96.9**	51.2
CROMA	44.3	-	-	-	-	48.4
SoftCon	54.5	-	-	-	-	<u>51.3</u>
AnySat	38.8	75.9	79.6	92.5	92.2	39.5
Galileo	40.4	-	-	-	-	39.5
SenPaMAE	40.7	59.9	79.5	89.5	78.3	39.3
DOFA	56.4	78.2	<u>80.4</u>	<u>92.8</u>	<u>96.3</u>	<u>51.3</u>
Panopticon	<u>59.3</u>	78.1	79.6	92.6	95.2	**52.6**

Table 5. GEO-Bench segmentation performance. We report mIoU in percentage.

s	2	4	8	13
MS_{acc}	81.2	85.3	83.6	81.2

Table 6. Ablation of non-overlapping spectral crop size.

References

[1] Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth c x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023. 5

[2] Jeremy Bentham. *Panopticon or the inspection house*. 1791. 2

[3] Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. Spectralearth: Training hyperspectral foundation models at scale. *arXiv preprint arXiv:2408.08447*, 2024. 3, 4

[4] Olivier Burggraaff. Biases from incorrect reflectance convolution. *Optics express*, 28(9):13801–13816, 2020. 4

[5] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 2

[6] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. *arXiv preprint arXiv:2407.03653*, 2024. 4

[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 *Advances in Neural Information Processing Systems*, 2022. 2

[8] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 7

[9] Michel Foucault. Panopticism. In *The information society reader*, pages 302–312. Routledge, 2020. 2

[10] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. 6

[11] Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Theres Kuester, André Hollstein, Godela Rossner, Christian Chlebek, et al. The enmap spaceborne imaging spectroscopy mission for earth observation. *Remote Sensing*, 7(7):8830–8857, 2015. 3

[12] Ethan King, Jaime Rodriguez, Diego Llanes, Timothy Doster, Tegan Emerson, and James Koch. Stars: Sensor-agnostic transformer architecture for remote sensing. *arXiv preprint arXiv:2411.05714*, 2024. 4

[13] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geobench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[14] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. *arXiv preprint arXiv:2405.02771*, 2024. 3

[15] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023. 4

[16] Jonathan Prexl and Michael Schmitt. Senpa-mae: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. *arXiv preprint arXiv:2408.11000*, 2024. 4

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 6

[18] Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. TorchGeo: deep learning with geospatial data. In *Proceedings of the 30th international conference on advances in geographic information systems*, pages 1–12, 2022. 2, 3

[19] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[20] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint arXiv:2310.18653*, 2023. 8

[21] 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. *arXiv preprint arXiv:2403.15356*, 2024. 4