Towards Geospatial Foundation Models via Continual Pretraining Supplementary Material

Matías Mendieta[1*] Boran Han[2] Xingjian Shi[3] Yi Zhu[3] Chen Chen[1]
[1] Center for Research in Computer Vision, University of Central Florida
[2] Amazon Web Services [3] Boson AI
matias.mendieta@ucf.edu boranhan@amazon.com xshiab@connect.ust.hk
yi@boson.ai chen.chen@crcv.ucf.edu

1. Overview

The supplementary material is organized into the following sections:

- Section 2: Training details for the pretraining stage and all downstream tasks.
- Section 3: Details on calculations of CO_2 impact.
- Section 4: Further analysis on the SpaceNet2 super-resolution task.

2. Training Details

We provide the training details for the various stages and tasks in our evaluation. Code, model weights, and GeoPile dataset are publicly available at https://github.com/mmendiet/GFM.

Change Detection: We modify the MMsegmentation [3] framework to conduct our change detection experiments. For OSCD, as the raw image size is large but the number of samples is very small, we tile the images into 192×192 pixels and train for 4000 iterations. We utilize the RGB bands for OSCD as in [8]. For DSFIN, we train for 10k iterations with image size 512×512. We employ an SGD optimizer with a learning rate of 0.01 and weight decay of 5.0e-4, and the default polynomial scheduler of [3].

Classification: On UC Merced, we train with a batch size of 1024 (128 per GPU) at image size 256×256. We train for 100 epochs with a base learning rate of 1.0e-4. We employ random flip, crop and standard Mixup [11] augmentation. Optimizer, weight decay, Mixup parameters, and other training settings are the same as in [10]. For BigEarthNet, we slightly upscale the original 120×120 images to 128×128 for ease of dimensional compatibility with the Swin transformer. We then employ the same training settings as with UC Merced.

Segmentation: We employ the MMsegmentation [3] framework to conduct our segmentation experiments. For both datasets, we train for 40k iterations with an image size of 512×512. All other training settings are the same as the default configuration in [3] for the respective backbones (Swin, ViT, ResNet50) and compatible decoders (UperNet [9] for transformers and Deeplabv3 [1] for ResNets).

Super-resolution: On the SpaceNet2 super-resolution tasks, we train with a batch size of 64 (16 per GPU) with input image size 160×160 and target size 640×640. We train for 100 epochs with a base learning rate of 1.25e-5. Optimizer, weight decay, and other training settings are the same as in [10], but with no random augmentations. We employ the standard decoder from [10] to produce the original input size from the encoder features, and then upscale using a convolution-based upsampling block based on the image reconstruction module for classic super-resolution employed in [6]. Detailed results for all downstream experiments and ablations from the main manuscript are provided in Table 2.

3. Training Time and Carbon Calculations

To calculate the CO_2 impact of training various models, we employ the ML CO_2 Impact estimator at https://mlco2.github.io/impact from [5]. The total impact is dependent on the hardware type, GPU provider, region, and total time used. Our pretraining experiments were conducted in the AWS US East (Ohio) region, which has a carbon efficiency of 0.57 kg eq. CO2 per kWh. For our GFM, just 93.3 V100 GPU hours are needed for training, resulting in a total carbon impact of 13.3 kg eq. CO2. This is significantly lower than the previous state-of-the-art geospatial model, SatMAE [2]. According to the reported carbon impact in their paper [2], SatMAE requires 768 V100 GPU hours and 109.44 kg eq. CO2 on the Google Cloud Platform us-central1 region, which has a carbon efficiency of 0.57 kg eq. CO2 per kWh (same as AWS US East Ohio). Therefore, GFM enables more than 8× reduc-

*Work done as an intern at Amazon Web Services

Table 1. SpaceNet2 super-resolution results with the residual connection.

Method	PSNR ↑	SSIM ↑
ViT (ImageNet-22k)[4]	22.548	0.629
SatMAE [2]	22.450	0.636
Swin (random) [7]	22.190	0.642
Swin (ImageNet-22k) [7]	22.918	0.640
GFM	**22.963**	**0.660**

tion in total training time and carbon impact in comparison to SatMAE.

4. Super-resolution with Residual Connection

In super-resolution tasks, a residual connection can be included from the input to the output stage [6]. We make this modification as well for both ViT and Swin, and present the results in Table 1. Interestingly, the Swin transformer benefits from this, while ViT does not. Nonetheless, in comparison to baselines, the conclusion is the same; SatMAE is not able to improve over its ImageNet-22k baseline, but GFM does.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 1

[2] 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. *arXiv preprint arXiv:2207.08051*, 2022. 1, 2

[3] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[4] 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. *CoRR*, abs/2010.11929, 2020. 2

[5] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 1

[6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 1, 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 2

[8] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *CoRR*, abs/2103.16607, 2021. 1

[9] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer, 2018. 1

[10] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *CoRR*, abs/2111.09886, 2021. 1

[11] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. 1

Table 2. Detailed downstream results for all experiments in the main manuscript. We abbreviate the following for horizontal space: UC Merced (UCM), BigEarthNet (BEN), WHU Aerial (WHU), Vaihingen (Vai), SpaceNet2 (SN2). † indicates vanilla continual pretraining.

Method	OSCD (F1)	DSFIN (F1)	UCM	BEN 10%	BEN 1%	WHU	Vai.	SN2 (PSNR)	SN2 (SSIM)
ImageNet-22k baseline	52.35	69.62	99.0	85.7	79.5	90.4	74.7	21.655	0.612
Sentinel-2	55.14	64.31	94.5	84.9	70.0	86.2	63.3	19.961	0.566
GeoPile	56.59	68.31	98.8	86.0	79.2	89.4	73.6	22.315	0.630
GeoPile†	57.10	66.88	98.7	86.2	79.3	90.0	74.6	22.566	0.638
GeoPile† (800ep)	57.52	66.23	98.8	86.3	79.3	90.1	75.1	22.626	0.645
Stage 1	56.20	69.79	98.1	85.8	78.3	89.0	73.3	22.153	0.626
Stage 2	58.97	68.27	96.9	86.1	79.0	89.4	72.2	22.409	0.625
Stage 4	60.31	68.97	98.3	86.1	80.8	89.8	73.0	22.495	0.638
Both Init.	58.01	69.77	98.5	85.8	77.2	90.1	74.1	22.930	0.669
w/o WHU-RSD46	58.79	69.25	98.3	86.1	80.6	89.7	72.9	22.510	0.632
w/o MLRSNet	60.01	69.21	98.8	86.1	80.5	89.9	72.9	22.409	0.633
w/o Resisc45	58.33	69.22	98.6	86.3	80.7	89.8	72.4	22.206	0.635
w/o PatternNet	59.00	70.37	98.3	86.3	80.5	89.8	71.9	22.293	0.629
w/o curated datasets	58.49	67.16	98.1	85.7	79.9	88.9	72.7	22.852	0.584
w/o NAIP	58.72	70.54	98.3	85.5	79.6	89.7	70.8	22.574	0.632
GFM	59.82	71.24	99.0	86.3	80.7	90.7	75.3	22.599	0.638