Parameter-Efficient Adaptation of Geospatial Foundation Models through Embedding Deflection: Supplementary material

A. Datasets Details

So2Sat is designed to detect environmental and urban dynamics and land cover changes in cities worldwide using remote sensing data. It combines Sentinel-1 (SAR) and Sentinel-2 (multispectral) imagery to track urban expansion and environmental changes. The dataset contains 17 bands, with imagery captured or upsampled at a 10-meter spatial resolution. The data is divided into 19,992 training images, 1,000 validation images, and 1,000 test images, each patch being 32x32 pixels. We selected, for our purpose, the Sentinel-2 bands.

Brick Kiln focuses on detecting and monitoring brick kilns, which are often associated with environmental pollution in rural areas. This dataset uses 13 bands from Sentinel-2 imagery. The images are 64x64 pixels in size, with a 10-meter resolution, and consist of around 15,000 samples for training and around 2,000 for validation and test.

ForestNet focuses on forest monitoring, aimed at identifying and classifying forested areas to support environmental conservation and management. It utilizes Landsat-8 imagery with 12 bands, including both multispectral and thermal bands, at a 15-meter resolution. The dataset has more than 8,000 patches, each 332x332 pixels.

MADOS is focused on detecting marine pollution, such as oil spills and debris, in oceanic environments. It contains Sentinel-2 imagery with bands 1-8A, 11, and 12 (8 bands total), covering various sea surface features from 174 scenes. The dataset is segmented into 2803 tiles, each measuring 240x240 pixels.

HLS BurnScars provides imagery for identifying burn scars, a critical task for monitoring the effects of wildfires on ecosystems and land. It uses Harmonized Landsat and Sentinel-2 (HLS) data with 6 bands: Blue (B02), Green (B03), Red (B04), NIR (B8A), SW1 (B11), and SW2 (B12). The imagery is available at a 30-meter spatial resolution, and the dataset consists of 804 scenes, each with a size of 512x512 pixels.

Each of these datasets plays a significant role in environmental monitoring using satellite imagery. They are designed to tackle a wide range of issues, including urban expansion, pollution, deforestation, forest management, and the impact of natural disasters like wildfires and oil spills. These datasets are essential for advancing our ability to monitor and manage the Earth's changing environments.

B. On DEFLECT implementation

Table 1. Ablation study about the computation of the spectral patch embeddings on MADOS

Model	w/ projection (default)	w/o projection
Scale-MAE	50.6	50.3
DINO-MC	51.6	48.5
Cross-Scale MAE	38.2	38.1

Reprojecting the features In the default setting of DEFLECT, the spectral pacth embeddings pass through a linear layer shared across attention blocks, such that the dimension of the spectral embeddings matches the dimension of the spatial embeddings. In some cases (depending on the number of multispectral channels), we can remove the projection layer, and only select the right number of statistics to have the right dimension. Table 1 shows the difference of test metrics on MADOS with and without the projection layer. It seems that having a projection layer can slightly improve the performance of DEFLECT.

Pixel-set encoding In our setting, the pixel-set encoding module randomly samples 10% of pixels within a patch (as we assumed that the spectral information was redundant within neighboring pixels). Table 2 shows the test metrics obtained on MADOS and BurnScars, by sampling 50% of the pixels, that do not significantly differ from a 10% sampling. These results confirm our hypothesis that there is strong spectral redundancy within a patch.

Standard attention VS Untangled Attention Fig. 1

Table 2. Ablation study about the pixel-set encoding module

	MADOS (mIoU)		BurnScars (IoU)	
Model	10%*	50%	10%*	50%
Scale-MAE	50.6	51.8	77.3	78.1
DINO-MC	51.6	50.3	75.6	75.0
Cross-Scale MAE	38.2	38.5	70.6	68.9

illustrates the differences between standard attention and our uAtt module.

Norm of the displacements Figure 2 suggests a general trend where larger changes in the norm of displacement between frozen and fine-tuned models lead to increased variability in test accuracy. This indicates that significant modifications to the displacement norm can negatively impact performance consistency across pretrained models. Notably, DEFLECT, which barely changes the norm of the displacement (after the first adapted layer, the norm can change with respect to the frozen pretrained GFM), exhibits lower standard deviation in test accuracy, demonstrating more robust results. This behavior highlights the potential advantage of controlling displacement norm variations and will be further investigated in future work.

C. Initialization strategies

HLSBurnScars To provide insights about the mechanisms of GFM adaptation to multispectral images, we investigated RGB weight initialization strategies, a topic often overlooked. The default method, widely used in tools like *timm*[3], involves repeating RGB weights across all bands, regardless of their physical meaning. An alternative, introduced by USat [1], initializes the RGB bands with pretrained weights while assigning random initialization to the others. For consistency with related works [2], we opted for the first method, here called *Repeat*, over the *RGB+random* approach. Tab. 3 summarizes the results on BurnScars, showing that *Repeat* generally achieved higher scores, though the performance varied significantly across PEFT methods. For instance, with Scale-MAE, *Repeat* yielded a mIoU of 80.1% under SLR (w.r.t 78.3% obtained with *RGB+random*) but dropped to 70.5% when using DINO-MC (w.r.t 73.2% with *RGB+random*). This variability suggests that while *Repeat* can offer slight advantages, neither approach provides robust, consistent results across methods. DEFLECT, that circumvents the choice of an initialization strategy, may thus provide a more reliable framework, avoiding the need for a sensitive selection of hyperparameters. Further results are detailed in the Supplementary Material.

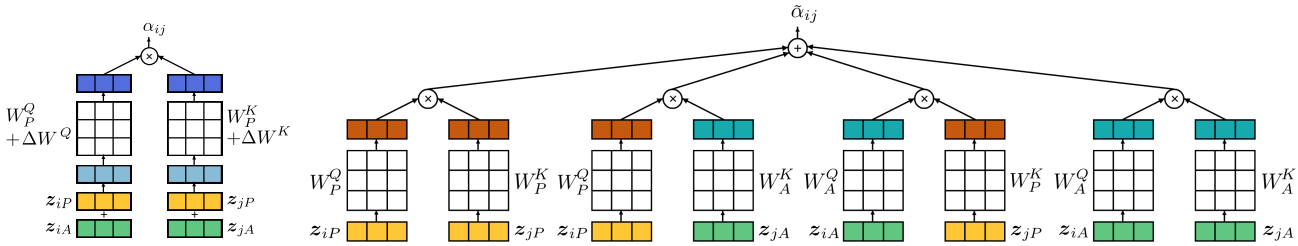
MADOS Table 4 reveals that for MADOS, the

Table 3. Performance metrics on BurnScars for two initialization strategies across different PEFT methods, showing similar but inconsistent results. This variability suggests that alternative approaches, such as DEFLECT, which do not depend on specific initialization schemes, may offer more stable and reliable performance in geospatial tasks.

Model	Tuning Strategy	Repeat	RGB+ random
Scale-MAE	Finetuning (Oracle)	79.1	75.5
	Frozen	76.2	68.9
	BitFit	76.0	72.0
	SLR	80.1	78.3
DINO-MC	Finetuning (Oracle)	76.5	76.9
	Frozen	70.4	66.9
	BitFit	69.8	66.6
	SLR	70.5	73.2
Cross-Scale MAE	Finetuning (Oracle)	78.1	75.8
	Frozen	68.0	59.9
	BitFit	67.9	68.5
	SLR	63.7	68.2

"RGB+random" initialization seems to perform slightly better across most methods. This trend further supports our hypothesis that different initialization strategies are not consistently reliable across different tasks. For example, "RGB+random" leads to higher mIoU scores in Scale-MAE (e.g. 53.1% vs. 47.0% for finetuning) and DINO-MC (e.g. 64.26% vs. 61.6% for finetuning). However, when we look at BurnScars (as shown in Table 5 in the main paper), the "Repeat" initialization yields better results on average, further emphasizing the variability of these initialization strategies.

This inconsistency confirms our key point: relying on specific initialization strategies is not ideal for stable, robust performance. Instead, DEFLECT, which operates independently of initialization schemes, proves to be a more reliable approach across diverse tasks, as it yields more stable results.



(a) Low-rank adaptation with standard self-attention

(b) DEFLECT adaptation with untangled attention

Figure 1. Illustration of standard attention and our untangled attention module in the context of PEFT.

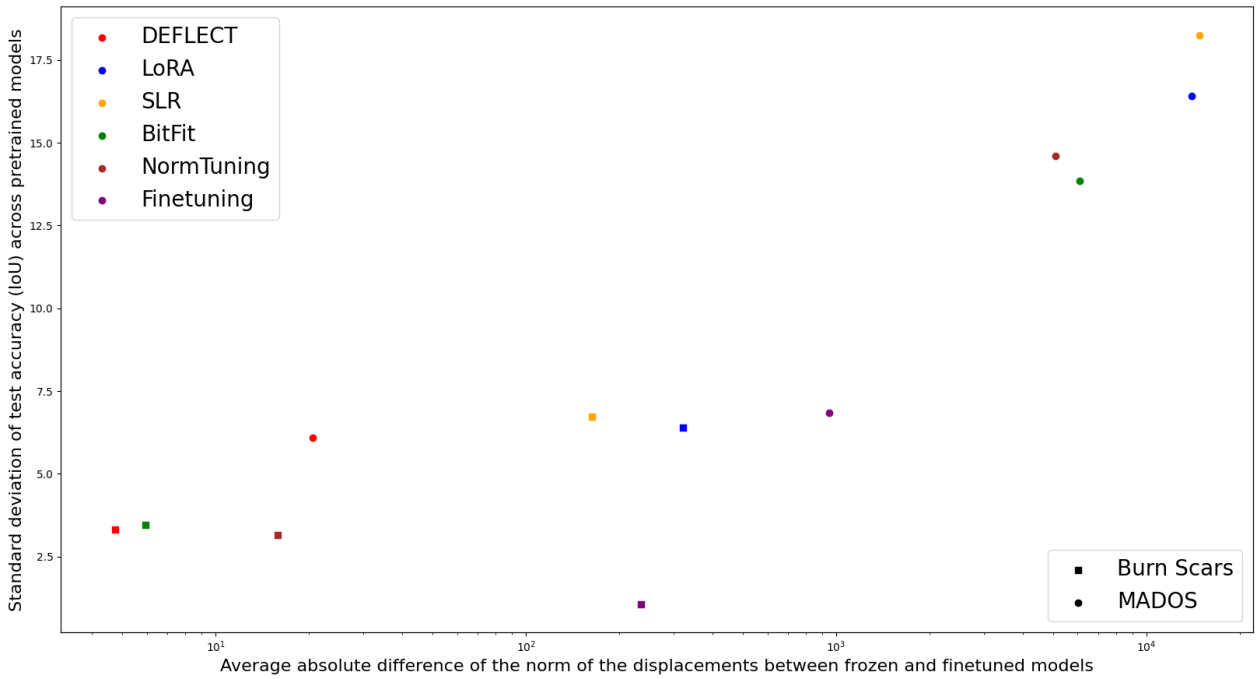


Figure 2. Standard deviation across models as a function of the average absolute difference of norm displacement.

Table 4. Performance metrics on MADOS for two initialization strategies across different PEFT methods. The table highlights the slight superiority of the RGB+random initialization, though results are inconsistent across datasets. This further supports our argument that reliance on initialization strategies can lead to unstable performance. In contrast, DEFLECT, which does not depend on specific initialization, demonstrates more reliable and stable results.

Model	Tuning Strategy	Repeat	RGB+ random
Scale-MAE	Finetuning (Oracle)	47.0	53.1
	Frozen	36.0	44.4
	BitFit	19.4	48.1
	SLR	46.5	46.9
DINO-MC	Finetuning (Oracle)	61.6	64.3
	Frozen	51.8	53.9
	BitFit	53.0	54.1
	SLR	3.5	5.8
Cross-Scale MAE	Finetuning (Oracle)	47.2	50.1
	Frozen	41.4	43.5
	BitFit	40.3	39.8
	SLR	35.6	36.1

References

[1] Jeremy Irvin, Lucas Tao, Joanne Zhou, Yuntao Ma, Langston Nashold, Benjamin Liu, and Andrew Y. Ng. Usat: A unified self-supervised encoder for multi-sensor satellite imagery, 2023. 2

[2] Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27841–27851, June 2024. 2

[3] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 2