GeoFoundation Model-Driven Semantic Segmentation of Wildfires

Authors: Esha Penmatsa, Kaushik Pillalamarri, Dr. Raju Vatsavai

Introduction: Over the past decade, forest wildfires have intensified globally, burning over 3.5 million km² annually, causing billions in economic losses, and triggering severe ecological disturbances, including biomass loss, soil erosion, and degradation of air and water quality. As a result, rapid, scalable burned area mapping is critical for emergency response, carbon accounting, and post-fire rehabilitation. While remote sensing enables synoptic monitoring, traditional operational products rely on spectral indices (e.g., dNBR) that are highly sensitive to atmospheric conditions and vegetation phenology. Supervised deep learning offers a powerful alternative for capturing nonlinear spectral-spatial patterns. Still, a severe scarcity of manually annotated fire boundaries leaves conventional models prone to overfitting across diverse ecoregions. Geospatial foundation models, which are pretrained on vast, unlabeled Earth observation corpora via self-supervised learning, offer the promise to bridge this gap by providing universal features that generalize well with limited labeled data through representations, transfer learning, and fine-tuning [8]. However, their relative efficacy on high-resolution, multi-band benchmarks such as CaBuAr, particularly under spectral heterogeneity and sensor-specific artifacts, remain largely uncharacterized.

State-of-the-art and Challenges: Early burned-area segmentation efforts have predominantly relied on U-Net and its variants [1]. While these architectures achieve competitive performance on standard benchmarks like the HLS BurnScar dataset (30m resolution, 6 bands) [6], they rely heavily on supervised training from scratch. Because they are trained end-to-end from random initialization, their performance degrades sharply when applied to out-of-distribution fire types (e.g., fast-moving grass fires vs. slow crown fires) or to sensors with different spectral response functions, severely limiting their geographic adaptability.

To address these limitations, geospatial foundation models, such as Prithvi EO 2.0 (NASA-IBM) [2], Galileo (Radiant Earth) [3], CROMA [4], and SMARTIES [7], have introduced robust, transferable representations using self-supervised pre-training. Although these models demonstrate strong zero-shot and few-shot capabilities on the BurnScar dataset [6], that benchmark represents a relatively controlled setting with consistent illumination and moderate burn scars.

In contrast, the California Burned Areas (CaBuAr) dataset presents a markedly more challenging benchmark that remains relatively unexplored. It comprises 20m Sentinel-2 Level-2A imagery featuring all 12 spectral bands (including coastal aerosol, SWIR, and cirrus) across diverse California ecosystems ranging from chaparral to coniferous forests [5]. CaBuAr presents severe class imbalances (where burned pixels often constitute less than 5% of a scene) alongside complex physical confounding factors. These include temporal misalignment between pre- and post-fire acquisitions due to cloud cover, resolution mismatches during ancillary data

fusion, and the inherent spectral ambiguity between burned soil and shadowed vegetation. Critically, no prior work has systematically benchmarked multiple foundation models on CaBuAr, nor explored whether their learned representations offer advantages for complex downstream segmentation.

Proposed Approach & Experimental Framework

In this work, we present a systematic comparative study of three geospatial foundation model backbones for binary burned area segmentation: **Prithvi EO 2.0** (a masked autoencoder trained on HLS time-series), **SMARTIES** (spectrum-aware masked autoencoder trained across heterogeneous sensors such as optical, radar, and sub-meter RGB), and **CROMA** (a contrastive learner on Sentinel-1/2 paired data).

Baseline Linear-Probe

Our experimental design strictly isolates representation quality from downstream task adaptation by employing a linear-probe protocol on paired pre- and post-fire imagery from the CaBuAr dataset. Each foundation model encoder is kept completely frozen. For the decoder we try two different methods, one is a lightweight, the first is a lightweight linear-probing head: two 1×1 convolutions (compressing the concatenated pre-/post-fire patch embeddings to a 256-channel bottleneck, then to a single logit map) separated by a GELU non-linearity, followed by bilinear upsampling to the input resolution. The second is a UPerNet[9]-style decoder that instead draws on four intermediate transformer layers spanning shallow to deep depths, resamples each layer's tokens to a distinct spatial scale to synthesize a four-level feature pyramid, aggregates global context at the coarsest level via a pyramid pooling module, and fuses all levels top-down through lateral convolutions before a final prediction layer. This controlled setup guarantees a fair, compute-efficient comparison of each model's intrinsic feature discriminability devoid of any fine-tuning.

For Prithvi, we also had to use a spectral adapter to feed all 12 bands of data to the model, as it is natively trained on a 6-band reflectance version of HLS data. To keep uniformity in the input provided, we have chosen to use all 12 bands of the Sentinel-2 multispectral data as provided by the CaBuAr dataset, and will run experiments in the future in native settings of the models chosen as part of an extended benchmark.

Training Configuration

We benchmark all configurations with the same setting, a frozen encoder with the decoder attached, whose weights are updated.

All experiments are conducted using a train-test split provided by the dataset, where all the splits in the data can be used for training except the ‘chabud’ split. But we further split the train data into train and val in an 80/20 ratio. To aggressively counteract the severe class imbalances

inherent in the CaBuAr dataset, the training pipeline incorporates an equally weighted sum of Dice loss and binary cross-entropy.

Experimental Results & Discussion (needs update)

Our experiments yield two principal findings regarding the performance of geospatial foundation models for burned area segmentation:

> **Multi-scale probing is advantageous over just final layer embedding:** The frozen Prithvi EO 2.0 encoder achieves a test-set F1 score of 0.7529 and a mean Intersection-over-Union (mIoU) of 0.7418, significantly outperforming SMARTIES but less than the CROMA model with a test set F! of 0.7728 and mIoU of 0.7579. CROMA outperforms the models on almost all the parameters except the precision, but is closer to the best-performing model. The performance of the models using the UperNet decoder is clearly visible as for each frozen encoder, its UperNet decoder variant outperforms the linear probe decoder variant, motivating the multi-scale early fusion for the next set of experiments.

The experiment results are summarized in the following table

Model	Decoder	Val Prec.	Val Rec.	Val F1	Val mIoU	Test Prec.	Test Rec.	Test F1	Test mIoU
Prithvi EO 2.0	UperNet	0.7405	0.8674	0.7989	0.8123	**0.972**	0.6113	0.7529	0.7418
Prithvi EO 2.0	Simple LP	0.7384	0.8677	0.7979	0.8114	0.9658	0.6169	0.7529	0.7418
SMARTIES	UperNet	0.7521	0.8654	0.8048	0.8172	0.9669	0.4612	0.6245	0.6484
SMARTIES	Simple LP	0.7315	0.8119	0.7696	0.7903	0.9572	0.4856	0.6443	0.6611
CROMA	UperNet	**0.7582**	**0.8868**	**0.8174**	**0.8272**	0.9594	**0.647**	**0.7728**	**0.7579**
CROMA	Simple LP	0.7422	0.7642	0.7531	0.7789	0.9674	0.5315	0.6861	0.6907
earlyfusion	LP					.9478	.5475	0.6941	.5315
late fusion	indv decoders					.9530	.5973	.7343	.5802

Conclusion & Future Outlook (needs update)

Our preliminary work showcases a principled way of using multiple geospatial foundation models for wildfire segmentation on the challenging CaBuAr dataset. CROMA models show promising improvements over other foundation models, due to a strong relationship with the data on which the models were trained. We are working on building a multi-model embedding fusion to exploit synergistic features from these models. Following the UperNet-style decoder, we can extract embeddings from selected layers of each frozen encoder and align via bilinear

interpolation to a common spatial grid. We are considering early fusion, multi-scale early fusion, and late fusion strategies. We will report the full results at the workshop and in the full paper.

To facilitate reproducible research and operational integration into active wildfire monitoring workflows, our code and trained segmentation head weights will be made publicly available.

References:

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[2] D. Szwarcman et al., "Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2412.02732

[3] G. Tseng, A. Fuller, M. Reil, H. Herzog, P. Beukema, F. Bastani, J. R. Green, E. Shelhamer, H. Kerner, and D. Rolnick, “Galileo: Learning global and local features in pretrained remote sensing models,” 2025. [Online]. Available: https://arxiv.org/abs/2502.09356

[4] A. Fuller, K. Millard, and J. R. Green, "CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 5506–5538, 2023. [Online]. Available: https://arxiv.org/abs/2311.00566

[5] D. R. Cambrin, L. Colomba, and P. Garza, "CaBuAr: California Burned Areas Dataset for Delineation," IEEE Geoscience and Remote Sensing Magazine, vol. 11, no. 3, pp. 106–113, 2023. [Online]. Available: https://arxiv.org/abs/2401.11519

[6] J. Jakubik, M. Schneider et al., “Prithvi-100m: A multi-temporal terrestrial foundation model for earth observation,” arXiv preprint arXiv:2310.18660, 2023, official reference for NASA-IBM Prithvi Foundation Model.

[7] G. Sumbul, C. Xu, E. Dalsasso, and D. Tuia, "SMARTIES: Spectrum-Aware Multi-Sensor Auto-Encoder for Remote Sensing Images," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2506.19585

[8] X. Xiao et al., "Foundation Models for Remote Sensing and Earth Observation: A Survey," arXiv:2410.16602, 2024. [Online]. Available: https://arxiv.org/abs/2410.16602

[9] Xiao, T., Liu, Y., Zhou, B., Jiang, Y. and Sun, J., 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 418-434