Exploring Cross-Modality Transfer Learning: Fine-Tuning Optical Pre-Trained Models on Synthetic Aperture Radar Image

Matthew “Mehdi” Hatami Goloujeh
PhD Student in Civil & Environmental Engineering
University of South Carolina
mhatami@sc.edu

1. Abstract

Transfer learning has become a pivotal technique for leveraging pre-trained models to address data scarcity challenges in deep learning applications. This study investigates the potential of transfer learning across different sensor modalities, focusing on fine-tuning a geospatial foundation model, **Prithvi**, pre-trained on optical imagery, for segmentation tasks on a small Synthetic Aperture Radar (SAR) dataset. We compare its performance to a UNet model trained from scratch on the same SAR dataset. Our primary hypothesis posits that the pre-trained optical data features in the Prithvi model can transfer effectively to SAR data, resulting in improved performance during fine-tuning. The experimental design evaluates both models using key metric, namely, Intersection over Union (IoU) and , to assess their ability to generalize despite the domain gap. Results from this study aim to provide insights into the transferability of features across modalities and the implications for data-scarce geospatial applications. This work advances understanding of cross-modality transfer learning and highlights the potential for integrating optical pre-training in SAR-based tasks.

2. Introduction

Flood mapping is a critical task in disaster management, providing essential information for response planning and mitigation. However, a significant challenge arises when attempting to train deep learning models for flood mapping in regions where data availability is limited. This project addresses the problem of flood mapping for such underrepresented regions by leveraging **transfer learning**. The approach involves fine-tuning models pre-trained on more data-rich regions or using different sensor modalities, enabling the development of accurate models for areas with scarce data resources.

A key challenge lies in fine-tuning the recently published geospatial foundation model, **Prithvi** [1], for a different data modality. The Prithvi model is pre-trained on optical imagery containing six bands (coastal, blue, green, red, near-infrared, and shortwave infrared 1). Both its encoder and decoder are tailored for optical data [2], making adaptation to Synthetic Aperture Radar (SAR) images nontrivial. The SAR dataset used in this study consists of 2-band images, representing conditions before and after flooding. A significant part of this work involved modifying the model's decoder to accommodate and fine-tune on the SAR dataset.

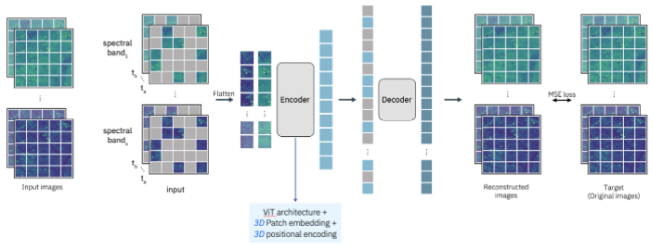

Figure 1: Prithvi model architecture

The importance of this work lies in the potential efficiency gains. If the Prithvi model's pre-trained features on optical data can effectively transfer to SAR data, there would be no need to retrain geospatial foundation models from scratch on SAR data, saving substantial computational resources. For example, training Prithvi originally required hundreds of GPUs over weeks of computation. By

contrast, fine-tuning for new modalities could dramatically reduce time and costs. The dataset used in this study was prepared specifically for this project. It comprises fewer than 150 SAR images, each of size 256x256 pixels, with two bands (pre- and post-flood) and corresponding flood masks. These masks were generated using a thresholding method via the **Google Earth Engine (GEE)**[3] platform, a cloud-based tool that facilitates access to large-scale geospatial datasets and computational resources for remote sensing analysis.

Preliminary results indicate the feasibility of cross-modality transfer learning for flood mapping, shedding light on the utility of leveraging pre-trained geospatial foundation models for data-scarce applications. This study not only advances understanding of transfer learning across modalities but also contributes to resource-efficient approaches in geospatial modeling.

3. Related Works

Flood mapping has widely utilized optical satellite data, such as Sentinel-2 and Landsat, due to its rich spectral information [4]. However, cloud cover during floods limits its usability, making Synthetic Aperture Radar (SAR) a critical alternative for its weather-independent capabilities. Deep learning with SAR data poses challenges due to its differing signal characteristics compared to optical data.

Geospatial foundation models, such as Prithvi, have demonstrated success in various geospatial tasks through pre-training on large-scale optical datasets. While these models enable resource-efficient fine-tuning, adapting them to SAR data remains underexplored [5]. Cross-modality transfer learning has shown promise in leveraging pre-trained features from optical data for SAR tasks, yet existing studies primarily focus on data fusion or smaller, task-specific models.

This study investigates fine-tuning the Prithvi model for SAR-based flood mapping using a small, custom SAR dataset. By exploring the transferability of optical pre-trained features, this work addresses a key gap in leveraging foundation models for modality-specific geospatial challenges.

4. Methods and Experiments

This study utilizes a custom-built dataset focusing on two distinct flood scenarios: the Valencia flood in October 2024 and the Mississippi flood in May 2019. The dataset preparation involved several systematic steps, leveraging Sentinel-1 SAR imagery and extensive pre- and post-processing to ensure high-quality, balanced data suitable for model fine-tuning and evaluation.

4.1. Data Collection

Sentinel-1 SAR images from before and after the floods were retrieved using the **Google Earth Engine (GEE)** platform, a cloud-based platform designed for planetary-scale geospatial analysis. GEE provides seamless access to a vast repository of satellite imagery, including Sentinel-1 data, and offers powerful tools for image processing and analysis. Its Python and JavaScript APIs enable users to perform tasks such as filtering by date, location, and acquisition mode, directly on the platform without the need for extensive local storage or processing power. In this study, GEE was instrumental in efficiently accessing SAR data and applying a series of preprocessing filters, such as temporal, spatial, and noise reduction filters, to ensure the raw images were optimized for subsequent analysis. This streamlined workflow significantly reduced the complexity and time required to prepare the dataset.

4.2. Preprocessing

To enhance image quality and relevance, the following filters were applied: **Temporal filter:** Ensured selection of images close to the flood

event dates. **Spatial filter:** Focused on the flood-affected regions. **Instrument mode filter:** Selected images captured in Interferometric Wide Swath mode for consistency. **Polarization filter:** Used VV and VH polarizations to highlight flood signals. **Orbit pass filter:** Choose images from descending orbits for uniformity. **Resolution filter:** Standardized the resolution to ensure comparability. **Noise reduction filter:** Applied to remove speckle and other artifacts inherent to SAR data.

4.3. Flood Mask Generation

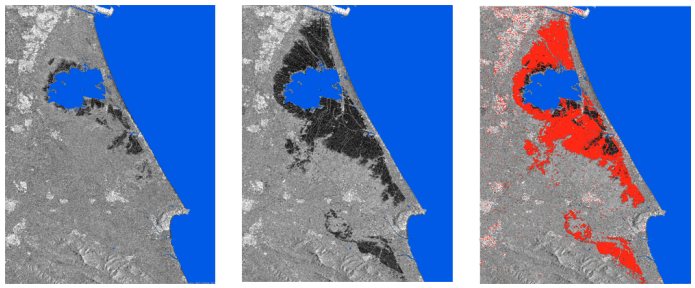

Figure 2 - from left to right, before flood, after flood, flood mask

Following the **SPIDER framework** (Synthetic Perturbation and Inundation Detection using Empirical Relationships), pre- and post-flood Sentinel-1 SAR images were analyzed to generate an initial flood mask. The SPIDER framework is a well-established methodology in remote sensing for detecting flood extents by exploiting differences in backscatter intensities between pre- and post-flood conditions. It involves applying thresholding techniques to identify regions with significant changes, indicative of flooding. This approach is particularly effective for SAR data, where water-covered areas typically exhibit low backscatter due to their smooth surface properties. After generating the initial flood mask, further refinement steps were undertaken to improve its accuracy and usability. **Permanent water bodies** were removed using auxiliary datasets, such as hydrological databases or water occurrence layers, to ensure only newly flooded areas were represented. Additionally, regions with a **slope greater than 5%** were excluded, as steep areas are less prone to flooding and could introduce false positives in the flood mask. The slope information was derived from Digital Elevation Models (DEMs) integrated with the dataset. These refinement steps ensured that the flood mask was both accurate and focused on relevant areas, improving its reliability for training segmentation models. This meticulous process resulted in high-quality flood masks tailored for each image pair, essential for building a robust and meaningful dataset for subsequent analysis.

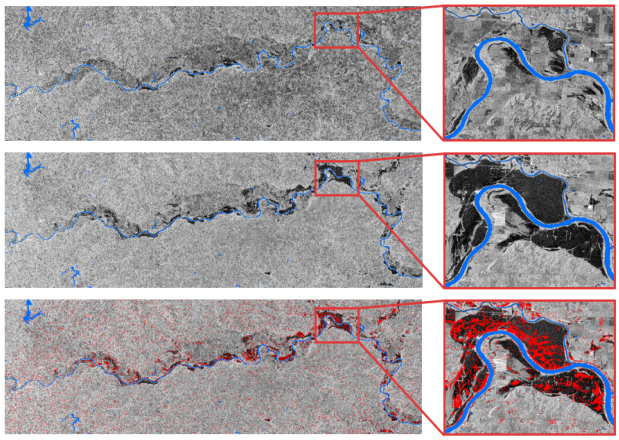

Figure 3 - from top to down, before flood, after flood, and flood mask

4.4. Post-Processing

Flood events are typically rare, leading to an **imbalanced dataset** where over 90% of the area remains unflooded. To address this, images with less than **10% flooded area** were excluded, resulting in a more balanced dataset. This step improved the training process by reducing the dominance of the unflooded class.

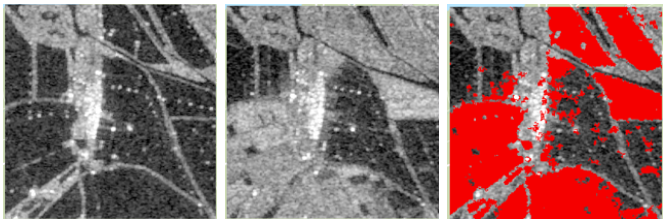

Figure 4 - from left to right, before flood, after flood, flood mask

4.5. Dataset Summary

The final dataset comprises 196 pairs of two-band images, where **Band 1** represents the pre-flood condition and **Band 2** represents the post-flood condition, each with a resolution of 256x256 pixels. Accompanying these image pairs are refined flood masks that delineate the flooded areas, offering sufficient detail for segmentation tasks. The dataset is organized into three folders: 156 images for training, 20 for validation, and 20 for testing. This dataset is uploaded on google drive and available at the link in the reference section [4]

4.6. Fine-Tuning the Prithvi Model

The Prithvi geospatial foundation model, pre-trained on large-scale optical imagery, was adapted for SAR-based flood mapping. The encoder from the original model, pre-trained on six optical bands (red, green, blue, short-wave infra-red-1, short-wave infra-red-2, and near infrared), was retained for feature extraction, while the decoder was replaced with a new architecture to process the 2-band SAR images (band 1 for before the flood, and band 2 for after flood). The modified model was fine-tuned using the training subset of the dataset, optimizing its parameters for the flood segmentation task. This fine-tuning step aimed to leverage the pre-trained encoder's feature representations to adapt to the new modality efficiently.

4.7. Training the UNet Network

For comparison, a **UNet network** was trained from scratch using the same SAR dataset. UNet is a well-established deep learning architecture widely used for image segmentation tasks due to its ability to capture fine-grained spatial details while maintaining contextual understanding. The network consists of a symmetric encoder-decoder structure with skip connections, which allow for the seamless integration of high-resolution features from the encoder to the decoder, enhancing segmentation accuracy. For this study, the UNet was adapted to process 2-band input images (pre- and post-flood conditions) and output corresponding flood masks. Training was conducted using the training subset of the dataset, with hyperparameters such as learning rate, batch size, and optimizer settings carefully tuned to minimize loss and maximize segmentation accuracy. This architecture served as a baseline to evaluate the effectiveness of transfer learning with the Prithvi model.

5. Results

The experiments were designed to evaluate the performance of two models, the Prithvi geospatial foundation model (fine-tuned) and the UNet network (trained from scratch), for flood mapping using a small SAR dataset. The results are summarized as follows:

5.1. Training and Validation for UNet

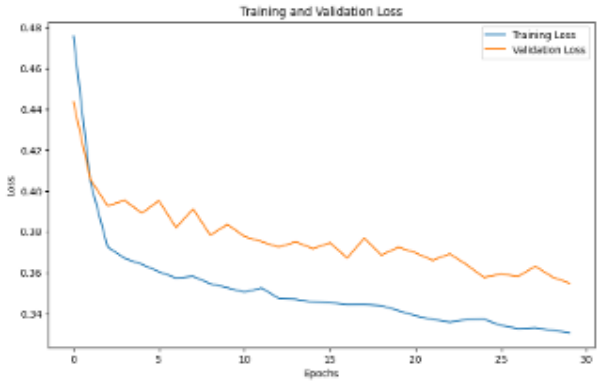

Figure 5: Training & validation loss for UNet

The training and validation loss curves for the UNet model over 30 epochs demonstrate a consistent downward trend for both lines, indicating that the model is effectively learning from the training data without overfitting to the validation set. This suggests that the UNet architecture is capable of learning meaningful patterns from the small SAR

dataset, albeit without the benefit of pre-trained features.

5.2. Fine-Tuning Prithvi

The fine-tuning process for the Prithvi model successfully adapted its pre-trained encoder to the SAR dataset by training a new decoder head. The output logs from the fine-tuning code indicate convergence during training, suggesting that the model was able to learn from the small dataset effectively by leveraging its pre-trained optical features.

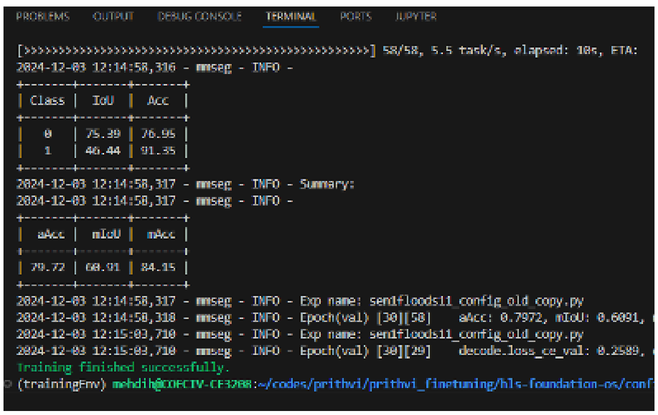

Figure 6: Prithvi model fine-tuning log

5.3. Intersection over Union (IoU) comparison

A direct comparison of the IoU scores between the two models highlights the superiority of the fine-tuned Prithvi model. The IoU for Prithvi significantly outperformed UNet, achieving a value of approximately **2.2**, compared to the UNet's **0.6**. This substantial difference confirms the hypothesis that the features learned by Prithvi during its pre-training phase on optical data are transferable and provide a strong foundation for learning SAR-based flood segmentation with minimal data. The results underscore the advantage of transfer learning for small datasets and cross-modality tasks.

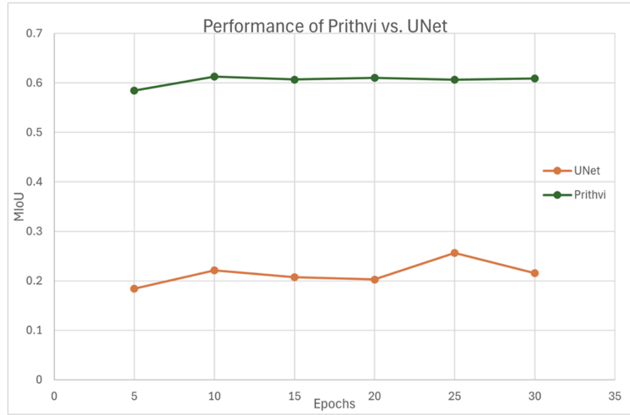

Figure 7: Comparing IoU for Prithvi vs. UNet over 30 epochs

6. Conclusion

This study investigated the use of cross-modality transfer learning for flood mapping with SAR data by comparing a fine-tuned Prithvi geospatial foundation model and a UNet network trained from scratch. The results demonstrate that the Prithvi model, pre-trained on optical imagery, significantly outperforms the UNet model on a very small SAR dataset, achieving a much higher IoU score. This confirms the hypothesis that features learned from optical data can be effectively transferred to SAR-based tasks, providing a resource-efficient solution for data-scarce scenarios.

The findings highlight the potential of leveraging pre-trained geospatial models for tasks requiring domain adaptation, particularly when training data is limited. Fine-tuning pre-trained models not only saves computational resources but also enhances performance compared to training from scratch. Future work could explore extending this approach to other geospatial tasks and sensor modalities, as well as further optimizing the integration of pre-trained features for cross-modality applications.

7. References

[1]https://huggingface.co/ibm-nasa-geospatial/Prithvi-EO-1.0-100M-sen1floods11/blob/main/sen1floods11-finetuning.png
[2]https://arxiv.org/pdf/2309.14500
[3]https://earthengine.google.com/
[4]https://idl.iscram.org/files/anastasiamoumtzidou/2020/2296_AnastasiaMoumtzidou_etal2020.pdf
[5]https://www.clarku.edu/centers/geospatial-analytics/projects/prithvi-foundation-model/