Toward a Global SAR Foundation Model for SAR-to-Spectral-Indices Mapping

Ziyi Song [1] **Noel Ivan Ulloa** [1] **Juntao Yang** [2] **Sang-Ho Yun** [1 3 4]

[1]*Earth Observatory of Singapore, Nanyang Technological University, Singapore* [2]*NVIDIA, Singapore* [3]*Asian School of Environment, Nanyang Technological University, Singapore* [4]*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore*. Correspondence to: Sang-Ho Yun sangho.yun@ntu.edu.sg.

1. Introduction

Foundation models pretrained on large-scale datasets have shown strong transferability across many domains. Such models have recently been developed for optical-only or multimodal satellite data, including large-scale optical models such as Prithvi [1] and multimodal models such as Clay [2], DOFA [3], and CROMA [4], while models tailored specifically to Synthetic Aperture Radar (SAR) remain underexplored. Moreover, most existing SAR-based models are still task and region specific, limiting their reusability and scalability.

SAR imagery provides all-weather, cloud-penetrating observations and captures physical surface properties, but it is noisy and difficult to interpret and annotate. In contrast, optical imagery like Sentinel-2 provides rich spectral information for land and vegetation monitoring, including widely used indices such as the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI), but it is frequently unavailable due to cloud cover. These complementary characteristics raise a practical question: can a globally pretrained SAR foundation model learn transferable representations for cross-modal tasks, such as estimating optical spectral information? We hypothesize that self-supervised pretraining on multi-temporal Sentinel-1 data can produce SAR representations that support SAR-to-spectral-indices mapping.

In this work, we develop a global SAR foundation model using a masked autoencoding (MAE) approach [5] with Vision Transformers (ViT) [6], pretrained on multi-temporal Sentinel-1 VV/VH time series. We evaluate transfer performance on a multi-band spectral indices regression task that predicts optical indices from SAR inputs. Our contributions include (1) a globally curated paired Sentinel-1 and Sentinel-2 dataset, (2) a self-supervised SAR foundation model trained on multi-temporal SAR data, and (3) an empirical study demonstrating the benefit of foundation pretraining for cross-modal SAR-to-spectral-indices regression.

2. Methodology

2.1 Dataset

We built a globally representative, multitemporal, and multimodal dataset from Sentinel-1 SAR and Sentinel-2 optical imagery. The data covers diverse geographic regions and climates, as illustrated in Figure 1, which shows the global spatial distribution of the GlobalPairs dataset (1M samples) and its overall geographic coverage. Each sample consists of five consecutive Sentinel-1 Ground Range Detected (GRD) frames (VV and VH polarizations) acquired at 12-day intervals, with the most recent SAR frame temporally aligned to a Sentinel-2 optical acquisition. During pretraining, only SAR data are used. The paired Sentinel-2 images are introduced later to supervise downstream tasks. For example, these paired optical images can be used to derive spectral indices for cross-modal finetuning in the SAR-to-spectral-indices task described in Section 2.3. This design supports both large-scale self-supervised pretraining and supervised downstream tasks using paired optical data.

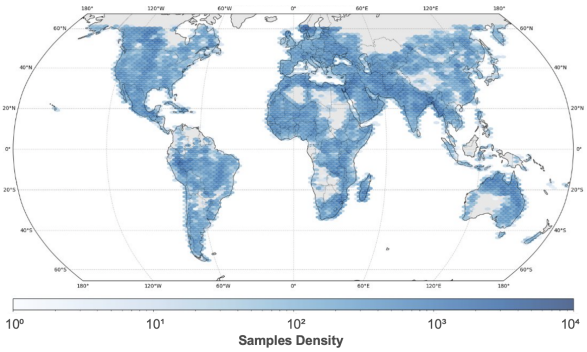


Fig. 1: Global spatial distribution of the GlobalPairs dataset (1M samples). Density is shown as the number of samples per unit area (logarithmic scale).

All models reported in this paper are pretrained on the 250 000-sample split unless otherwise specified. For scale tests, we prepared datasets of different sizes, including 250 000, 500 000, 1 million, and 2 million samples.

2.2 Model Architecture and Pretraining

The foundation model was trained using a MAE approach with a ViT backbone on multitemporal SAR data. During training, 75% of each input is randomly masked, and the encoder learns to reconstruct the full image by seeing only the remaining 25%. This learning objective encourages the encoder to capture spatial and temporal SAR patterns without relying on labels. Additional architectural details are provided in Appendix A. We ran pretraining on 250 000 samples for 1 000 epochs using 8 NVIDIA A100 GPUs.

2.3 Downstream Task: Multi-band Spectral Indices Regression

To evaluate the learned SAR representations and demonstrate their transferability, we fine-tuned the

pretrained ViT encoder with a neck and task-specific regression head for SAR-to-spectral-indices mapping using MSE loss. The model takes multi-temporal Sentinel-1 VV/VH inputs (t_0 to t_4) and predicts six spectral indices for the last timestep t_4: Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Modified Normalized Difference Water Index (MNDWI), Normalized Difference Moisture Index (NDMI), Normalized Burn Ratio (NBR), and Enhanced Vegetation Index (EVI). Target labels are derived from temporally paired Sentinel-2 acquisitions at t_4. The fine-tuning dataset consists of 1 600 training samples, with 85 held-out test samples. Fine-tuning was performed on a single NVIDIA H100 GPU and completed in approximately 12 minutes. We compare against a U-Net [7] baseline and a from-scratch variant to assess the benefit of pretraining.

3. Results and Discussion

3.1 Pretraining Evidence

To evaluate our pretrained SAR foundation model, we visualize its ability to reconstruct time-series Sentinel-1 imagery from 75% masked inputs. Using a ViT-Base encoder pretrained on 250 000 SAR samples over 1000 epochs, we tested reconstruction on a cropland scene with five temporal frames (t_0 to t_4). For each frame, we compare: (1) the original image, (2) the 25% visible masked input, (3) the model's reconstruction, and (4) an overlay combining the reconstruction and visible patches, as shown in Fig. A2. Visualization is rendered in QGIS, where SAR backscatter is mapped using false-color to highlight structural details and feature interpretation. The raw Sentinel-1 VV+VH grayscale input at t_0 is shown in Fig. A2.

Despite seeing only 25% of the input, the model is able to recover spatial features such as field boundaries, circular structures, and linear patterns. Stronger backscatter regions remain noticeably brighter, maintaining relative intensity. Even though random speckle frequently produces salt-and-pepper noise in SAR images, the model can still extract physical features beyond this noise. These results indicate that the model captures meaningful spatiotemporal patterns, supporting its use as a foundation for downstream tasks.

3.2 Multi-band Spectral Indices Regression

Table 1 reports performance on 85 test samples for each band. The model achieves the best performance on vegetation-related indices (NDVI: R^2=0.79, EVI: R^2=0.77) and water-related indices (NDWI: R^2=0.74), while moisture indices (NDMI: R^2=0.57) remain more challenging.

Figure A3 in Appendix A shows qualitative results on a representative test sample. Only NDVI and NDWI are visualized out of 6 bands as they best correspond to the dominant land cover types (forest and water). The model successfully recovers spatial patterns that distinguish vegetated areas from water bodies using SAR inputs alone, although boundary delineation remains imperfect and images appear slightly blurry.

Table 1: Per-index regression performance of the SAR-pretrained ViT model on 85 test samples.

Index	R^2 ↑	RMSE ↓
NDVI	0.79	0.094
NDWI	0.74	0.093
MNDWI	0.67	0.101
NDMI	0.57	0.101
NBR	0.66	0.112
EVI	0.77	0.104
Mean	**0.70**	**0.101**

Table 2: Comparison of different models. Metrics are reported as the unweighted mean across six spectral indices (per-index average).

Model	Mean R^2 ↑	Mean RMSE ↓
ViT (SAR pretrained)	**0.70**	**0.101**
U-Net (ImageNet pretrained)	0.59	0.118
U-Net (scratch)	0.56	0.123
ViT (scratch)	0.44	0.138

3.3 Effect of Pretraining and Baseline Comparison

To evaluate the benefit of SAR-specific pretraining, we compare a ViT-based regression model with and without ViT-MAE pretrained weights, and U-Net (ResNet-34 [8]) baselines with and without ImageNet pretraining.

As shown in Table 2, the SAR-pretrained ViT model achieves a mean per-index R^2 of 0.70 and RMSE of 0.101, compared to 0.44 and 0.138 when trained from scratch. This improvement indicates that self-supervised pretraining on large-scale SAR data produces transferable representations for cross-modal regression tasks.

Notably, the scratch-trained ViT (112M) underperforms the smaller U-Net models (25M), while the SAR-pretrained ViT outperforms all baselines, indicating that the observed gains are driven by domain-specific pretraining rather than model size or architecture.

4. Challenges and Future Work

Several challenges remain. SAR speckle noise may affect feature learning, and the current model produces predictions with limited boundary sharpness and patch artifacts. Future work includes exploring the effect of temporal sequence length on model performance, and extending the SAR foundation model to broader Earth observation tasks. In addition, we plan to evaluate the predicted spectral indices in downstream application scenarios, such as vegetation monitoring and surface water mapping, to assess their practical utility and inform further model development.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 3 Award MOE-MOET32022-0006. We acknowledge the institutional support from the Earth Observatory of Singapore (EOS), Nanyang Technological University (NTU). We thank the NTU High Performance Computing Center and the National Supercomputing Centre (NSCC), Singapore, for providing computational resources. We also acknowledge support from the NVIDIA Academic Hardware Grant Program for computing resources used in this work. The Sentinel-1 and Sentinel-2 data used in this study were provided by the European Space Agency (ESA) and accessed through Google Earth Engine. We also acknowledge the use of generative AI tools for checking language clarity while preparing this abstract.

References

[1] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Dario Oliveira, Naomi Simumba, Shraddha Singh, Linsong Chu, Devyani Lambhate, Karandeep Singh, Daiki Kimura, Hamed Alemohammad, Radoslav Ganchev, Korneli Sabo, Arlindo da Silva, Campbell Watson, Manil Maskey, and Raghu Ganti. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. Accepted at NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning.

[2] Clay Foundation. Clay Foundation Model. `https://github.com/Clay-foundation/model`, 2024. Accessed: 11.01.2025.

[3] 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 multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

[4] Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2311.00566.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. Available at `https://doi.org/10.48550/arXiv.2111.06377`.

[6] 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. *arXiv preprint arXiv:2010.11929*, 2020. Available at `https://doi.org/10.48550/arXiv.2010.11929`.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Appendix A. Model Architecture and Qualitative Results

This appendix provides additional details on model architecture and qualitative results.

Figure A1 illustrates the ViT-MAE architecture used for SAR foundation model pretraining.

Figure A2 shows reconstruction results on multi-temporal Sentinel-1 imagery, demonstrating the model's ability to recover spatial features from masked inputs.

Figure A3 presents qualitative results for spectral indices regression on a representative test sample.

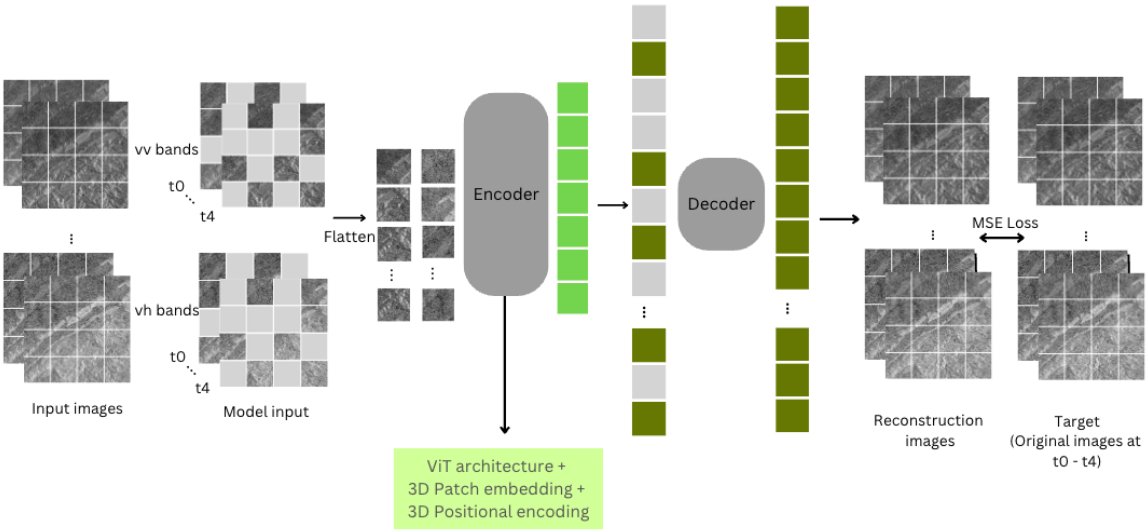
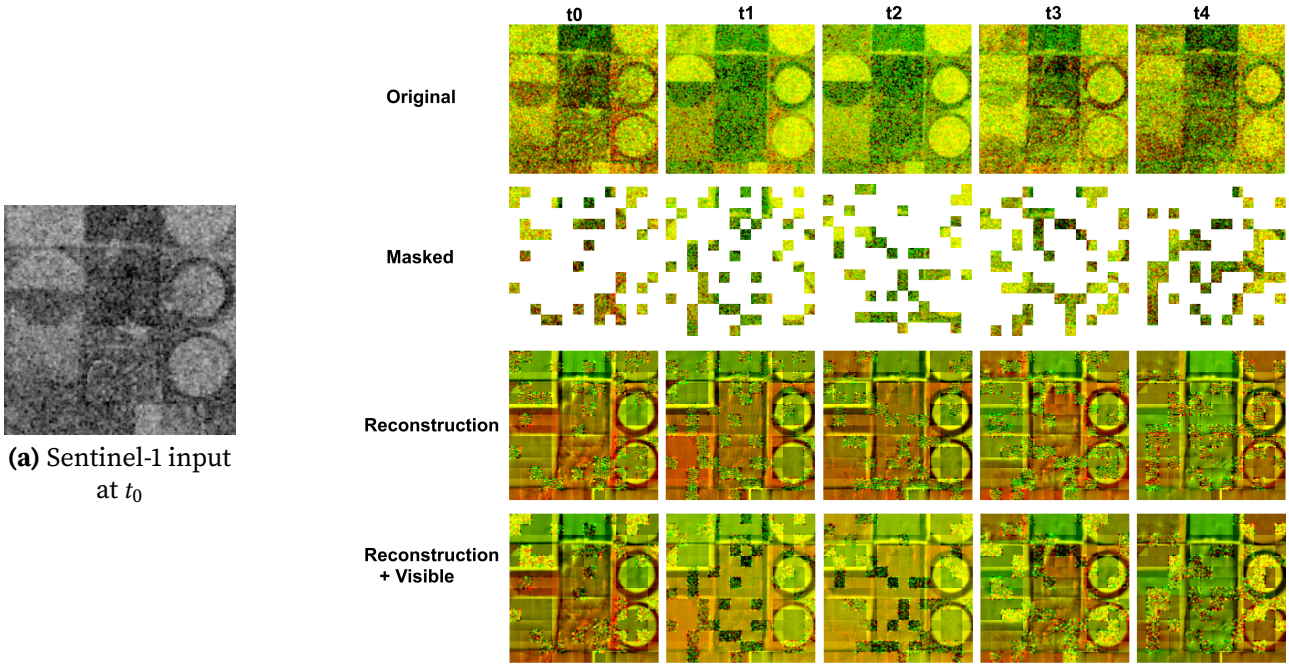


Fig. A1: Overview of the ViT-MAE architecture used for SAR foundation model pretraining.



(a) Sentinel-1 input at t_0

(b) Reconstruction results (t_0–t_4)

Fig. A2: Reconstruction results of multi-temporal Sentinel-1 imagery. (a) Original SAR input at t_0 (VV+VH grayscale). (b) Visualization of model reconstruction across time steps.

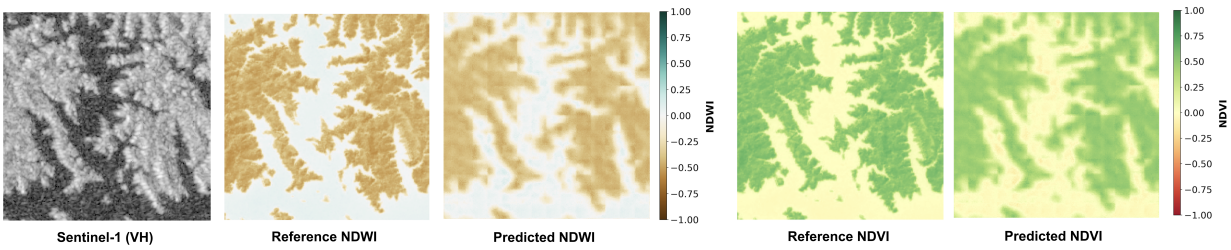


Fig. A3: Qualitative results on a test sample from the Thailand-Myanmar border region. From left to right: Sentinel-1 VH input (t_4), NDWI reference and prediction, NDVI reference and prediction.