Architectural Adaptation of Remote Sensing Foundational Models for Object Detection

Devansh Lodha
devansh.lodha@iitgn.ac.in

Rishabh Mondal
rishabh.mondal@iitgn.ac.in

Nipun Batra
nipun.batra@iitgn.ac.in

Sustainability Lab **IIT Gandhinagar**

1. Introduction

Problem: Manual satellite image analysis for environmental monitoring (e.g., brick kilns, industrial sites) is unscalable.

Challenge: State-of-the-art Vision Transformer (ViT) based RSFMs produce **1D token sequences**, which are incompatible with object detectors expecting **2D feature maps**.

Solution: We developed a **generalized adapter framework** to bridge this architectural gap, enabling object detection. This work benchmarks RSFMs, explores new feature engineering, and translates research into an actionable tool for stakeholders.

2. Generalized ViT-FPN Adapter

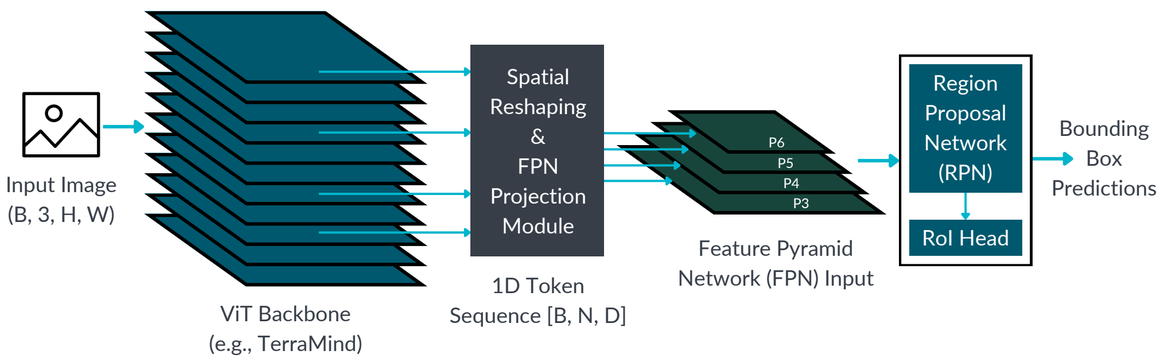

Fig 1. Generalized ViT-FPN Adapter Architecture. Our adapter resolves the architectural mismatch between ViT-based RSFMs and detection heads. It extracts token sequences from multiple transformer layers and reshapes them into a multi-scale feature pyramid (P3 to P6), providing the detector with both high-resolution spatial details and rich semantic context.

3. SOTA on SentinelKilnDB

NEURAL INFORMATION PROCESSING SYSTEMS 2025

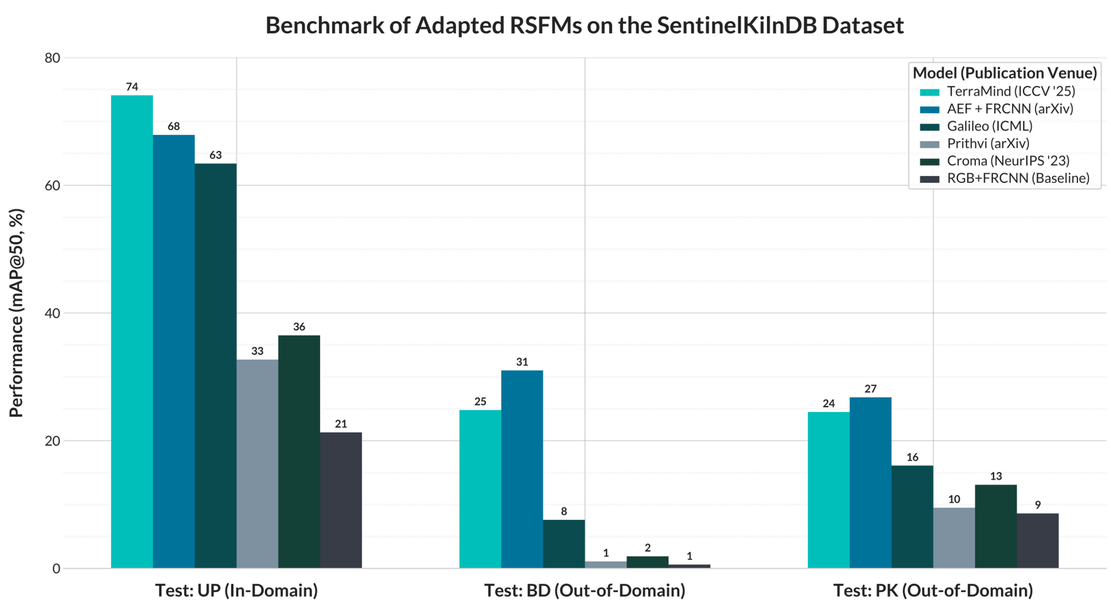

Fig 2. RSFM Generalization Performance. All models use a Faster R-CNN head and are trained on the Uttar Pradesh (UP) dataset. While adapted RSFMs significantly outperform the RGB baseline, performance degrades on out-of-domain data (BD, PK), motivating our work on domain adaptation.

We also benchmarked **AlphaEarth (AEF)** by treating its 64-D embeddings as a **64-channel semantic image**. As shown, this method is highly competitive, proving that semantic context can outperform raw pixel data.

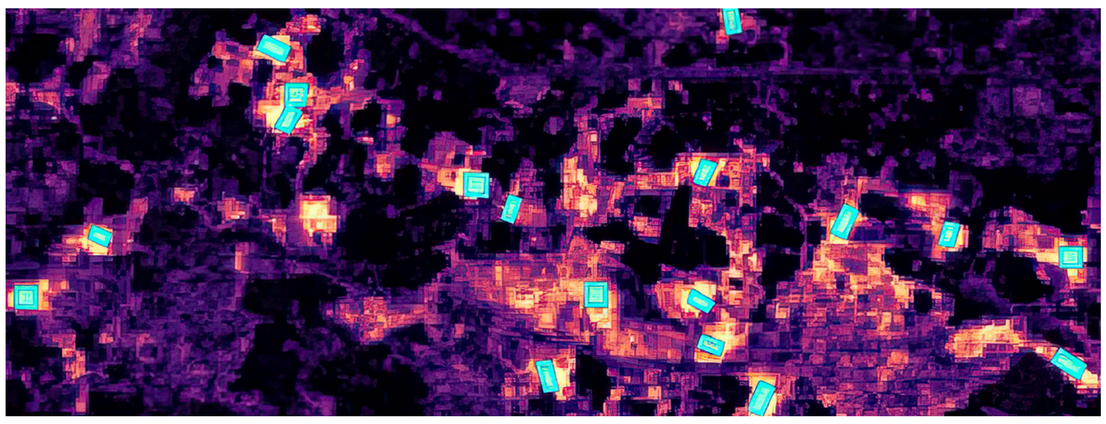

Fig 3. Application of Google Deepmind's AlphaEarth embeddings. A segmentation heatmap trained on embeddings reveals a distinct semantic signature for brick kilns.

4. Dual-Stream Prototypical Alignment for Unsupervised Domain Adaptation

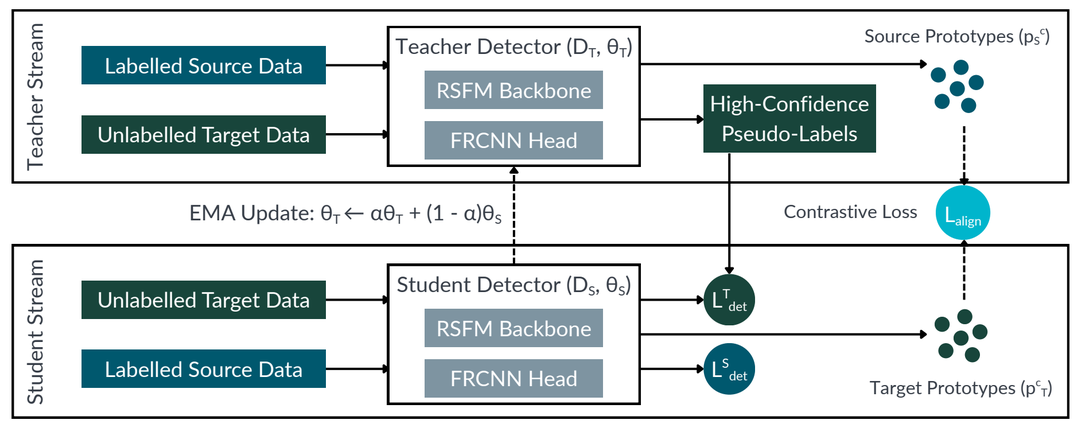

Fig 4. Proposed Dual-Stream UDA Architecture. The Student Detector (D_S) is trained on labeled source and unlabeled target data. The Teacher Detector (D_T) is a slow-moving average of the student (via EMA), providing stable pseudo-labels and source-domain prototypes (p_S^c). The core innovation is the Prototypical Alignment Loss (L_{align}), a contrastive loss that pulls the student's target-domain prototypes (p^c_T) towards the teacher's source-domain prototypes of the same class, while pushing them away from other classes, thus learning a domain-invariant feature space.

5. Real World Impact and Validation

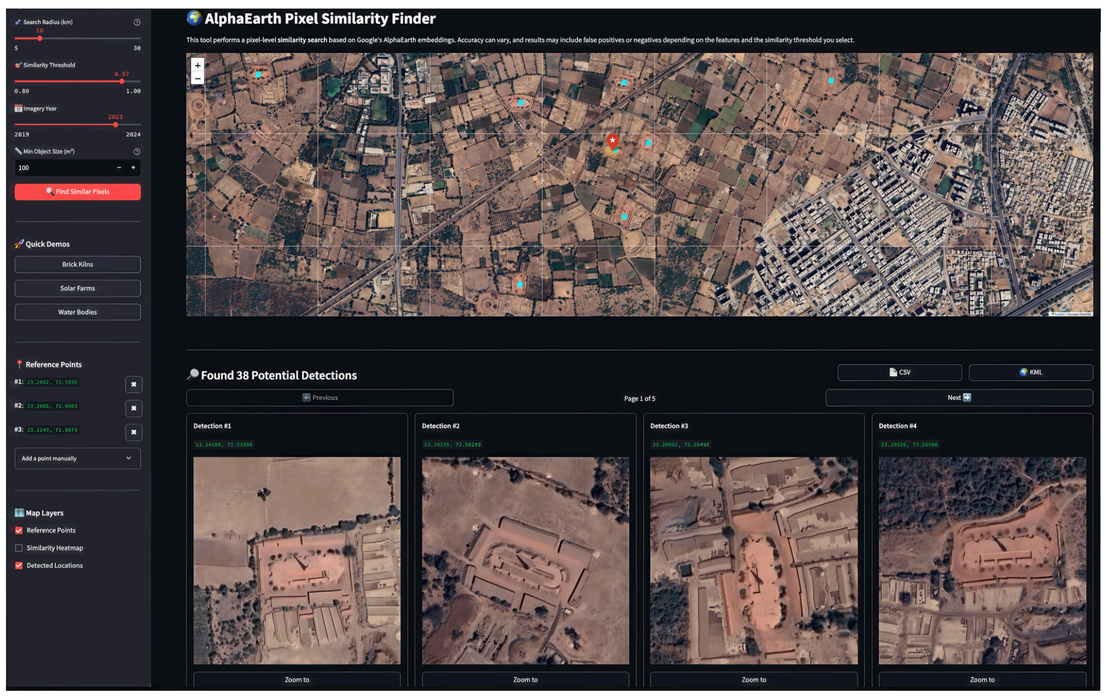

Fig 4. The Alpha-Earth Pixel-Similarity Finder. A deployed web application translating our research into an interactive tool for expert use.

SpatialThoughts

CEEW THE COUNCIL

Scan to use the app!

The application's architecture and scaling strategy were validated by **Spatial Thoughts** (Ujaval Gandhi, ex-Google Geo Data Strategist).

Evaluated by **Council on Energy, Environment and Water (CEEW)** for:

- Mapping small-scale industrial units for emissions tracking.
- Identifying agricultural burning sites and solar farms.

References

- Brown, C. F., et al. (2025). AlphaEarth Foundations: An embedding field model for accurate and efficient geospatial analysis. arXiv preprint.
- Chen, T., et al. (2020). A Simple Framework for Contrastive Learning of Visual Representations. ICML.
- Fuller, A., et al. (2023). CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. NeurIPS.
- Jakubik, J., et al. (2023). Foundation models for generalist geospatial artificial intelligence. arXiv preprint. [Prithvi]
- Jakubik, J., et al. (2025). TerraMind: Large-Scale Generative Multimodality for Earth Observation. ICCV.
- Liu, Y., et al. (2021). Unbiased Teacher for Semi-Supervised Object Detection. ICLR.
- Mondal, R., et al. (2025). SENTINELKILNDB: A Large-Scale Dataset and Benchmark for OBB Brick Kiln Detection... NeurIPS.
- Ren, S., et al. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NeurIPS.
- Snell, J., et al. (2017). Prototypical Networks for Few-shot Learning. NeurIPS.
- Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS.
- Tseng, G., et al. (2025). Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. ICML.