WORKSHOP ON EARTH OBSERVATION AI FOUNDATION MODELS

Authors: Prof. Rafael Kargren (ESA/ SmartSAT CRC), Prof. Tat-Jun Chin (University of Adelaide), Associate Prof. Bihan Wen (NTU)

Executive Summary

This white paper synthesizes findings from a collaborative workshop on Earth Observation AI Foundation Models held as a Special Session of 2025 IAC in Sydney, Australia.

1. Introduction

1.1 Background

Satellite-based remote sensing generates vast quantities of imagery across multiple modalities—optical, synthetic aperture radar (SAR), hyperspectral, and others. However, the development of robust AI systems for analysing this data faces unique challenges compared to terrestrial computer vision applications. These include:

- Limited annotated training data, particularly for specialized modalities
- Extreme variability in imaging conditions and environmental contexts
- Computational and power constraints of space-qualified hardware
- Radiation-induced bit flips and hardware faults in orbital environments
- Need for rapid adaptation to novel categories and changing conditions

Traditional approaches requiring extensive labelled datasets and task-specific model architectures prove inadequate for the dynamic requirements of space-based AI systems. This workshop explored physics-driven approaches that incorporate domain knowledge from imaging systems and atmospheric physics into AI architectures, enabling more efficient learning and robust generalization.

1.2 Workshop Participants and Focus Areas

Over 70 participants from arounds the globe participated in this IAC Special Session representing mix of academia, research institutes, space agencies and commercial companies. The workshop was organised and hosted by Prof. Rafael Kargren of ESA / SmartSAT CRC

Keynote presenters included Prof. Tat-Jun Chin of University of Adelaide's AI for Space Group and Associate Prof. Bihan Wen of Nanyang Technological University's, exploring the intersection of artificial intelligence, computational imaging, and space-based remote sensing. The research addresses three critical challenges: data scarcity in specialized imaging modalities, AI generalization across novel scenarios, and reliable edge computing deployment in space environments.

Second part of workshop was hands-on discussion in groups focusing to engage the audience to discuss tangible use cases for AI EO foundation models and barrier identification and solution mapping. At the end of the workshop each table presented shortly their findings and suggestions.

2. Challenge 1: Data Scarcity and Synthetic Data Generation

2.1 The Problem of Limited Training Data

While optical satellite imagery benefits from decades of archived data and commercial availability (Sentinel-2, Landsat, etc.), specialized modalities face severe data scarcity:

- SAR imagery: Complex acquisition requiring specialized satellites, weather-independent but expensive
- Polarimetric SAR (PolSAR): Even more limited, with annotation costs exceeding $50 per image
- Hyperspectral imagery: Limited satellite platforms, specialized ground truth requirements
- High-resolution commercial imagery: Licensing restrictions and limited geographic coverage

Traditional approaches requiring 10,000+ annotated examples per class prove impractical for many remote sensing applications, particularly for rare events (disasters, maritime incidents) or novel infrastructure types.

2.2 Customized SAR Image Simulation (University of Adelaide)

The University of Adelaide team developed SN6-SAROPT, a novel pipeline for generating synthetic PolSAR images with automatic annotation:

Methodology:

1. 3D Model Extraction from OpenStreetMap (OSM): Retrieve building footprints, heights, and semantic labels
2. Single-Look Complex (SLC) Simulation: Generate SAR backscatter using ray-tracing and electromagnetic scattering models
3. Polarimetric Synthesis: Create full PolSAR imagery (HH, HV, VV polarizations)

Technical Implementation:

- Integration with Blender 3D engine for geometric modelling
- Physical SAR simulation engine accounting for:
    - Incidence angle dependencies
    - Multi-bounce scattering from building facades
    - Volumetric scattering from vegetation
    - Surface roughness characteristics

Dataset Characteristics:

- 724 paired SAR-RGB images
- Multiple geographic locations and building types
- Automatic semantic segmentation masks
- Various imaging geometries and resolutions

Validation Results: The synthetic data demonstrated effectiveness in training object detection models, with performance comparable to models trained on limited real SAR data. Cross-validation on real SAR datasets showed mIoU improvements of 8-12% when synthetic data augmented small real datasets.

2.3 AI-Driven EO to SAR Translation (Nanyang Technological University)

NTU researchers developed generative AI approaches for translating abundant optical imagery into synthetic SAR:

Architecture Components:

1. Conditional Generative Adversarial Networks (GANs):
    - Generator network learning optical-to-SAR mapping
    - Discriminator enforcing realistic SAR texture and speckle characteristics
    - Cycle consistency losses ensuring semantic preservation
2. Physics-Informed Constraints:
    - SAR-specific loss functions capturing speckle statistics
    - Geometric consistency across polarization channels
    - Shadow and layover direction constraints

Comparative Performance: Testing against baseline methods (Pix2Pix, CycleGAN, NICEGAN) showed the proposed approach achieved:

- Superior texture realism (perceptual similarity scores 15-20% higher)
- Better preservation of geometric structure (structural similarity index +0.12)
- More accurate classification transfer (optical labels applicable to synthetic SAR with 85%+ accuracy)

Application to Object Detection: Generated synthetic SAR enabled training robust detection models for 15 object categories with automatically transferred annotations from optical imagery:

- Plane, ship, vehicle, storage tank, bridge, harbor, swimming pool, tennis court, basketball court, roundabout, baseball diamond, ground track field, small vehicle, helicopter, landslide

Detection performance on real SAR imagery achieved mAP scores 18-25% higher than models trained only on limited real SAR data.

2.4 Integration of Multiple Data Sources

Both approaches demonstrated the value of leveraging complementary data modalities:

Physical Simulation Advantages:

- Explicit control over imaging parameters
- Perfect ground truth generation
- Ability to simulate rare conditions

AI Translation Advantages:

- Learns statistical relationships from real data
- Captures subtle texture and radiometric characteristics
- Adapts to actual sensor characteristics

Hybrid Strategies: Combining physics-based simulation with AI-driven refinement offers optimal results—physics models provide structural accuracy while AI components add realistic texture and sensor-specific characteristics.

3. Challenge 2: Foundation Models for Zero-Shot Generalization

3.1 Limitations of Task-Specific Models

Traditional remote sensing AI systems require separate models for each:

- Geographic region (training on European data may not generalize to Asian landscapes)
- Imaging condition (seasonal variations, atmospheric effects)
- Sensor type (different satellites have unique radiometric characteristics)
- Task (detection, segmentation, classification each require distinct architectures)

This proliferation of specialized models creates:

- Prohibitive training data requirements (10,000+ examples per task-region-sensor combination)

- Deployment complexity (multiple large models consuming limited satellite storage)
- Inflexibility (inability to respond to novel scenarios without retraining)

3.2 Foundation Model Approach

Foundation models, pre-trained on massive datasets and adaptable to diverse tasks, offer a paradigm shift for remote sensing AI. The workshop explored two complementary aspects:

3.2.1 Compact Geospatial Foundation Models (University of Adelaide)

Base Architecture: Starting from IBM/NASA's Prithvi-EO-2.0-300M—a Vision Transformer (ViT) backbone with masked autoencoder (MAE) pre-training on Harmonised Landsat-Sentinel-2 (HLS) imagery.

Compactification Strategy: To enable onboard deployment, the Adelaide team applied dual MAE knowledge distillation:

1. Teacher Model: Original Prithvi-300M (frozen)
    - 1024-dimensional patch embeddings
    - 12 transformer layers
    - 303M parameters
    - ~1157 MB model size (FP32)
2. Student Models: Reduced variants
    - Prithvi-512: 512-dimensional embeddings → 76M parameters, 290 MB
    - Prithvi-256: 256-dimensional embeddings → 19M parameters, 73 MB

Knowledge Distillation Process:

- Student trained to match teacher's output embeddings, not original data reconstruction
- Preserves semantic understanding while reducing capacity
- Progressive distillation maintaining intermediate layer alignments

Performance Analysis:

Downstream Task Evaluation: Five tasks across diverse datasets

Task	Dataset	Prithvi-300M	Prithvi-256	Degradation
Cloud Classification	Sentinel-2 Cloud Mask	97.2%	95.2%	-2.0%
Cloud Segmentation	Sentinel-2 Cloud Mask	90.1 mIoU	88.2 mIoU	-1.9 mIoU
Flood Detection	Sen1Floods11	93.4% F1	91.8% F1	-1.6% F1
Landslide Detection	Landslide4Sense	87.6 mIoU	85.3 mIoU	-2.3 mIoU
Above-Ground Biomass	ForestNet	48.2 RMSE	51.7 RMSE	+3.5 RMSE

The compact Prithvi-256 model maintains 92-98% of original performance while reducing model size by 94% and inference time by 65%.

Specialized Task Heads: Each downstream task requires only small decoder networks:

- Cloud classification: 2-layer MLP (133k-526k parameters, 0.51-2.01 MB)
- Segmentation tasks: UNet decoder (726k-972k parameters, 2.77-3.71 MB)
- Regression tasks: UPerNet decoder (2.87M-3.82M parameters, 10.95-14.56 MB)

Task heads can be uploaded separately (10-15 MB each), enabling mission flexibility without re-uploading the entire foundation model encoder.

3.2.2 Zero-Shot Detection and Segmentation (Nanyang Technological University)

Challenges in Remote Sensing Zero-Shot Learning:

1. Inter-class Similarity:
    - Basketball courts vs. tennis courts differ primarily in markings
    - Ships vs. large vehicles may appear similar in low-resolution imagery
    - Contextual information becomes critical for disambiguation
2. Intra-class Variance:
    - Ships range from small fishing vessels to massive container ships
    - Buildings vary enormously in size, shape, and material
    - Perspective distortions at different satellite viewing angles
3. Domain Gap:
    - Remote sensing imagery has top-down perspective unlike natural images
    - Object scales vastly differ (buildings span 10-100+ pixels)
    - Multispectral channels beyond RGB not present in vision pre-training datasets

Architecture: KMA-ZSIS (Knowledge-grounded Mask Attention Zero-Shot Instance Segmentation)

Component 1: Multi-Modal Feature Alignment

- CLIP (Contrastive Language-Image Pre-training) image encoder extracts visual features
- CLIP text encoder processes class descriptions
- Knowledge-grounded Mask Attention (KMA) module aligns spatial features with class embeddings
- Attention mechanism highlights class-relevant regions

Component 2: Class-Agnostic Mask Generation

- Proposal generator identifies potential object regions without class constraints
- Mask generator creates precise segmentation boundaries
- Operates on visual features alone—no semantic labels required during mask prediction

Component 3: Zero-Shot Classification

- Weighted classification head combines:
    - CLIP semantic similarity scores (text-image alignment)
    - Geometric context from regional features
    - Cache bank of prototypical embeddings from training

Context-Aware Extension: RS-CLIP (Region-Aware Semantic Context Integration)

Remote sensing objects rarely appear in isolation—context provides crucial disambiguation cues:

Regional Context Encoding:

1. Scene Context Branch:
    - Captures overall environment type (port, airport, urban, agricultural)
    - Processes full image with dilated convolutions
    - Generates scene-level embeddings
2. Patch Embeddings:
    - Local feature representation of proposal regions
    - Maintains fine-grained spatial details
3. Region-Aware Integration:
    - Transformer encoder fuses patch and scene embeddings
    - Adaptive region formation groups semantically similar patches
    - Learns which context scales matter for each object type

Global Context Adaptation:

- Instance-Class Similarity Index (ICSI): measures alignment of proposal features with class embeddings
- Adaptive fusion weights determined by context relevance
- Linear projection layers adapt CLIP features to remote sensing domain

Experimental Validation:

Dataset: SIOR (Salient Instance Object in Remote sensing)

- 800 images, 2,000+ instances
- 5 base classes (training): ship, vehicle, bridge, storage tank, harbor
- 4 novel classes (zero-shot testing): small car, large vehicle, airplane, wind turbine

Results:

Method	Novel mAP	Base mAP	Harmonic Mean
Baseline ViT	42.3	68.5	52.1
CLIP-based	56.8	72.1	63.5
KMA-ZSIS (Ours)	68.4	76.3	72.2
RS-CLIP (Ours)	73.9	79.1	76.4

Context integration improved novel class detection by 5.5 mAP points, with particularly strong gains for ambiguous categories:

- Small car vs. vehicle: +12% disambiguation accuracy
- Ship in harbor vs. ship at sea: +8% accuracy improvement

Generalization to Novel Datasets:

The trained models transferred to unseen datasets without fine-tuning:

- FAST (Fine-grained Object Categories): 65.3% mAP on 4 novel categories
- DIOR (Object Detection in Optical Remote Sensing): 58.7% mAP on 8 novel categories
- SODA-A (Small Object Detection): 48.2% mAP (challenging due to small object scales)

Qualitative Analysis:

Visualization of attention maps revealed:

- RS-CLIP correctly attends to surrounding infrastructure for ships (docks, cranes suggest cargo vessels)
- Runway context helps distinguish civilian vs. military aircraft
- Roads and parking patterns differentiate vehicle types
- Failure cases primarily occur when context is ambiguous or absent (isolated objects in featureless backgrounds)

3.3 Data Efficiency Comparison

Few-Shot Learning Experiments:

Comparing zero-shot foundation models vs. fully-supervised ViT trained from scratch:

Cloud Classification Task:

Training Data Fraction	ViT from Scratch	Prithvi-256 GeoFM	Prithvi-256 (Pretrained Head)
100% (full dataset)	94.2% Acc	95.0% Acc	95.8% Acc
50%	91.3% Acc	94.1% Acc	94.6% Acc
25%	85.7% Acc	91.8% Acc	92.3% Acc
0% (zero-shot)	N/A	N/A	87.2% Acc

Key Findings:

- GeoFM with randomly initialized task head matches full training performance with only 50% of labeled data
- Pre-trained task heads enable reasonable zero-shot performance (87% vs. 96% best-case)

- Gains most pronounced in limited data regimes—foundation models reduce labeling needs by 50-75%

Flood Detection Task:

Training Data Fraction	ViT from Scratch	Prithvi-256 GeoFM
100%	89.2% F1 (water)	93.1% F1 (water)
50%	82.5% F1	90.8% F1
25%	74.3% F1	86.9% F1

Even with 75% reduction in training data, the foundation model approach achieves F1 scores within 6 points of full supervision, while the from-scratch model degrades by 15 points.

3.4 Implications for Mission Design

Foundation models fundamentally change satellite AI deployment strategies:

Traditional Approach:

- Pre-mission: Identify specific tasks, collect training data, train specialized models
- Launch: Deploy fixed models
- Operations: Execute predetermined tasks only
- Limitations: Cannot adapt to unexpected scenarios, novel object categories, or changing mission priorities

Foundation Model Approach:

- Pre-mission: Deploy general-purpose encoder (one-time ~300 MB upload)
- Launch: Basic task heads for initial objectives
- Operations:
    - Upload new task heads as priorities evolve (10-20 MB per task)
    - Zero-shot inference on unexpected targets without updates
    - Few-shot adaptation with minimal labeled examples from ground station
- Advantages: Flexible mission evolution, rapid response to emerging needs, reduced pre-launch uncertainty

4. Challenge 3: Reliable Edge Computing in Space

4.1 Constraints of Onboard AI Processing

Space-qualified computing hardware faces extreme constraints compared to terrestrial AI infrastructure:

Hardware Limitations:

- Processing power: 10-100x slower than commercial GPUs

    - Rad-hardened chips trail commercial tech by ~5 years
    - Clock speeds: hundreds of MHz vs. GHz terrestrial processors
- Memory: 4-16 GB RAM (vs. 80+ GB for modern GPUs)
- Storage: 10-100 GB solid-state (vs. TB-scale terrestrial systems)
- Power: 5-10W for AI accelerator (vs. 300-700W for datacenter GPUs)
- Thermal management: passive cooling only, strict temperature ranges

Radiation Effects:

- Single-Event Upsets (SEUs): bit flips in memory and computation
    - Low-Earth Orbit: ~10^-7 upsets per bit per day
    - Geostationary/beyond: ~10^-5 upsets per bit per day (South Atlantic Anomaly, solar events)
- Cumulative radiation damage: gradual performance degradation
- Mitigation: Error correction, redundancy, rad-hardened components (expensive, lower performance)

Communication Constraints:

- Downlink bandwidth: 10-100 Mbps during ground station passes
- Pass duration: 5-15 minutes per ground station per orbit
- Total daily downlink: ~1-10 GB (vs. TB of raw imagery collected)
- Latency: Minutes to hours between capture and ground processing

Mission Justification for Onboard AI: Despite constraints, onboard processing enables:

- Prioritized downlinking (transmit only high-value imagery)
- Real-time alerts (disasters, military activity, maritime incidents)
- Autonomous tasking (retarget sensors based on detected events)
- Reduced ground processing burden (send analytics, not raw pixels)

4.2 Kanyini Mission: Flight-Representative Testing

Mission Overview:

- Launch: August 2024
- Orbit: Sun-synchronous LEO, 500km altitude
- Payload: Hyperspectral imager (400-1000nm, 120 spectral bands)
- Platform: 6U CubeSat
- Partners: SmartSat CRC, University of South Australia, University of Adelaide

Onboard Computing:

- Intel Myriad 2 Vision Processing Unit (VPU)
    - 12 SHAVE (Streaming Hybrid Architecture Vector Engine) cores
    - 4GB LPDDR3 memory
    - Neural compute engine optimized for CNN inference
    - Power: 1-2W typical, 2.5W peak

    - FP16 computation

Ground Testing Infrastructure:

Hardware Emulator:

- Development board matching flight hardware specs
- Controlled power/thermal monitoring
- Allows rapid software iteration before flight

HyperScout-2 Engineering Model:

- Flight-representative payload interface
- Simulates actual data flows and timing
- Validates end-to-end data processing pipeline

Model Deployment Process:

1. Model Optimization:
    - Training: PyTorch on GPU workstations (FP32)
    - Quantization: Convert to FP16 (halves memory, minimal accuracy loss)
    - Compilation: OpenVINO toolkit generates optimized VPU executables
    - Pruning: Remove redundant parameters (10-15% size reduction with <1% accuracy loss)
2. Flight Software Integration:
    - Real-time operating system (RTOS) with deterministic scheduling
    - Memory management ensuring no swapping/paging
    - Watchdog timers reset on inference timeout/failure
    - Error correction codes (ECC) on critical data structures
3. Validation Testing:
    - Functional: Verify output correctness across input ranges
    - Performance: Measure inference time, power consumption, memory usage
    - Fault injection: Simulate bit flips, verify error handling
    - Thermal: Test across operational temperature range (-40°C to +85°C)

4.3 Performance Results

Downstream Task Execution on Flight Hardware:

Cloud Detection (Tile-based Classification):

- Task: Classify 224×224 tiles as cloudy/clear
- Model: Prithvi-256 encoder + 2-layer MLP head
- Inference time: 5.36 seconds per tile
- Memory usage: 636 MB peak
- Power: 5.76W peak, 4.95W average
- Energy: 0.72 Wh per tile

- Accuracy (engineering model testing): 97.1% (vs. 97.2% GPU baseline)

Cloud Segmentation:

- Task: Pixel-wise cloud mask generation
- Model: Prithvi-256 encoder + UNet decoder
- Inference time: 5.50 seconds per tile
- Memory usage: 633 MB peak
- Power: 5.77W peak, 4.94W average
- Energy: 0.76 Wh per tile
- mIoU (engineering model): 90.1 (vs. 90.1 GPU baseline)

Flood Detection Segmentation:

- Task: Identify water vs. land vs. cloud
- Model: Prithvi-256 encoder + UPerNet decoder
- Inference time: 5.68 seconds per tile
- Memory usage: 634 MB peak
- Power: 6.20W peak, 4.97W average
- Energy: 0.79 Wh per tile
- F1 (water, engineering model): 91.6% (vs. 91.6% GPU baseline)

Key Observations:

FP32 vs. FP16 Accuracy Impact:

- Classification tasks: <0.5% accuracy degradation
- Segmentation tasks: <0.3 mIoU degradation
- Regression tasks: 2-3% RMSE increase (most sensitive to quantization)
- Conclusion: FP16 acceptable for most remote sensing tasks

Power and Thermal:

- Peak power (~6W) well within VPU limits
- Typical power (~5W) sustained indefinitely without thermal issues
- Engineering model testing: Stable operation at +60°C ambient
- Flight operations plan: Inference during eclipse (cooler thermal environment)

Throughput Analysis:

- ~650 tiles per hour (continuous inference, no downlink)
- Typical hyperspectral capture: 5000×5000 pixels = 484 tiles
- Full scene processing: ~45 minutes
- Realistic duty cycle (25%, accounting for capture/downlink): 2-3 full scenes per day
- Compared to ground processing: Results available within 1 hour vs. 6-12 hours for full downlink and processing

Failure Mode Testing:

- Watchdog timeout recovery: <2 seconds reset and restart
- Memory exhaustion handling: Graceful degradation (skip tiles if memory constrained)
- Simulated bit flips: Error detection in 98.7% of cases, corrected or flagged

4.4 ISS Demonstration: IMAGIN-e Payload

Mission Context:

- Platform: International Space Station, Bartolomeo external platform
- Deployed: Q4 2024 (currently operational)
- Partners: Thales Alenia Space, Microsoft, ESA
- Duration: 1-year demonstration mission

Edge Computing Hardware:

- 16 ARM Cortex-A72 cores (quad-core configuration, 4 units)
- 16 GB RAM total
- 10 GB usable storage for AI models
- Power budget: 25W for computing payload
- Thermal: Active cooling available (ISS advantage)

Microsoft Azure Orbital Space SDK:

- Containerized AI model deployment
- Ground-based model updates via S-band uplink
- Telemetry and logging infrastructure
- Over-the-air (OTA) software updates

Foundation Model Deployment:

Baseline Configuration:

- Prithvi-300M encoder (full-size foundation model)
- Multiple task heads deployed simultaneously:
    - Cloud detection (2 MB)
    - Flood segmentation (15 MB)
    - Agricultural monitoring (12 MB)
    - Urban change detection (18 MB)
- Total footprint: ~1.3 GB (encoder + 4 task heads)

Operational Workflow:

1. Capture imagery (simulated downlinked RGB Sentinel-2 data)
2. Route to appropriate task head based on ground command or autonomous scene classification

3. Run inference
4. Downlink compact analytics (segmentation masks, class labels, confidence scores) ~1% of raw image size
5. Ground operators review results, request specific raw imagery if needed

Results (Preliminary):

- Successful deployment and operation for 3+ months
- Average inference time: 8-12 seconds per 512×512 tile (varies by task head complexity)
- 99.2% inference success rate (failures mainly due to memory constraints on complex scenes)
- Demonstrated OTA task head upload: New "ship detection" head uploaded successfully (14 MB, 8-minute transfer)
- Power consumption: 18-22W during inference (within budget)

Radiation Effects:

- ISS in LEO, ~400 km altitude, partially shielded by Earth's magnetosphere
- Observed SEU rate: ~2-3 per day (logged in telemetry)
- Impact: 1 inference failure over 90 days attributed to uncorrected bit flip
- ARM Cortex-A72 includes ECC on caches, significant hardware resilience

4.5 Robustness to Radiation-Induced Faults

Bit Flip Simulation Studies:

To assess model vulnerability, researchers inject random bit flips in deployed models:

Methodology:

- Randomly flip N bits in model weights (simulating SEU without correction)
- Re-run inference on test set
- Measure accuracy degradation vs. number of flips

Results for Prithvi-256 (19M parameters, ~76 MB):

Number of Bit Flips	Cloud Classification Acc	Cloud Segmentation mIoU	Flood Detection F1
0 (baseline)	95.2%	88.2%	91.6%
1	95.1%	88.1%	91.5%
10	94.8%	87.9%	91.2%
100	93.4%	86.5%	89.8%
1000	87.2%	78.3%	82.1%
10000	52.6%	41.2%	47.3%

Key Findings:

- Single-bit flips: Minimal impact (<0.5% degradation)
- Moderate corruption (100 flips, ~0.0005% of parameters): <2% degradation
- Severe corruption (1000 flips): Significant but often mission-acceptable degradation
- Catastrophic failure threshold: ~10,000 flips (0.05% of parameters)

Redundancy and Error Correction:

Hardware ECC:

- Memory ECC: Single-error correction, double-error detection (SECDED)
- Typical overhead: 12.5% (8 parity bits per 64 data bits)
- Reduces effective memory by ~10% but prevents most SEUs

Software Strategies:

- Periodic model checksum verification
- Re-upload weights if corruption detected (via stored hash)
- Ensembling: Run inference with multiple weight snapshots, vote on results (3x computational cost but high resilience)

Graceful Degradation:

- Prioritize protection of early layers (capture low-level features, less redundancy)
- Later layers more redundant (higher-level semantic features robust to small perturbations)
- Critical layers (final classification head) protected with strongest ECC

Comparison: Foundation Models vs. Specialized Models:

Hypothesis: Foundation models' redundancy (trained on massive diverse data) provides inherent robustness.

Experiment: Compare bit flip resilience of:

1. Prithvi-256 (foundation model, 19M params)
2. Task-specific ViT trained from scratch (15M params, comparable capacity)

Number of Bit Flips	Foundation Model Acc	Specialized Model Acc
100	93.4%	91.8%
500	90.1%	85.6%
1000	87.2%	76.3%

Foundation models maintain higher accuracy under identical fault conditions, likely due to:

- Broader feature representations (less reliance on specific weights)
- Pre-training on diverse data creates robust embeddings

- Over-parameterization relative to downstream task provides redundancy

5. Multimodal AI for Natural Disaster Monitoring

5.1 Beyond Computer Vision

Limitations of Imagery-Only Approaches:

Satellite imagery provides rich spatial information but lacks critical context for disaster monitoring:

- Temporal lag: Optical satellites require clear skies (clouds obscure disasters)
- Limited physics: Images capture reflected light, not underlying processes
- No predictive capability: Imagery is reactive, not anticipatory

Value of Complementary Data Modalities:

Atmospheric Data:

- Temperature profiles: Identify heat anomalies (wildfires, volcanic activity)
- Humidity and precipitation: Predict flood risk, track storm development
- Wind patterns: Model fire spread, pollution dispersal
- Pressure systems: Early warning for tropical cyclones

Sensor Data:

- Seismic measurements: Earthquake detection, aftershock prediction
- Ocean buoys: Tsunami wave height, sea surface temperature
- River gauges: Real-time flood levels
- Weather stations: Ground truth for atmospheric models

SAR Imagery:

- All-weather capability: Penetrates clouds, operates day/night
- Coherent change detection: Millimeter-scale surface deformation
- Flood extent mapping: Water vs. land discrimination regardless of visibility

5.2 Unified Multimodal Foundation Model (NTU Research)

Architecture Design:

Modality-Specific Encoders:

1. SAR Multispectral Encoder:
    - Input: PolSAR data (HH, HV, VV polarizations) + derived products (coherence, entropy)
    - Architecture: ViT backbone, specialized positional encoding for SAR geometry

    - Output: 256-dimensional embedding per spatial patch
2. Atmospheric Data Encoder:
    - Input: 3D gridded atmosphere (temperature, humidity, wind at multiple pressure levels)
    - Architecture: 3D convolutional network
    - Output: 256-dimensional embedding per spatial location
3. Optical Image Encoder:
    - Input: Multispectral satellite imagery (RGB + NIR + SWIR bands)
    - Architecture: Shared ViT backbone with SAR encoder (transfer learning)
    - Output: 256-dimensional embedding per spatial patch

Modality Fusion:

- Self-attention over concatenated embeddings
- Learnable modality-specific position encodings
- Cross-attention between modalities captures complementary information
- Output: Unified 256-dimensional multimodal embedding

Self-Supervised Pre-Training:

Contrastive Learning Objective:

- Positive pairs: Different modalities observing same geographic location at same time
- Negative pairs: Same modality, different locations or times
- Loss: Maximize agreement between positive pairs, minimize for negatives
- Encourages learning modality-invariant representations

Masked Reconstruction:

- Randomly mask patches in each modality
- Predict masked content from other modalities
- Teaches model to leverage complementary information

5.3 Application: Flood Prediction and Mapping

Task Definition: Given multimodal inputs (SAR, optical, atmospheric data), predict:

1. Flood probability map (next 24 hours)
2. Flood extent map (current flooded areas)

Dataset Construction:

- Historical flood events: 150 floods across Southeast Asia (2018-2023)
- Co-located observations:
    - Sentinel-1 SAR (pre- and post-flood)
    - Sentinel-2 optical (when available)
    - ERA5 atmospheric reanalysis (temperature, precipitation, humidity)

- Ground truth: Copernicus Emergency Management Service flood maps

Model Configuration:

- Encoder: Unified multimodal foundation model (pre-trained on 500K global observations)
- Task head: UNet-style decoder for dense prediction
- Training: Fine-tune on 120 flood events, test on 30 held-out events

Baseline Comparisons:

Method	Modalities	Flood Detection F1	Flood Prediction AUC
SAR-only	SAR	85.3%	72.1%
Optical-only	Optical	78.6% (clouds limit data)	68.4%
Atmosphere-only	Temperature, precip, humidity	N/A	74.8%
SAR + Optical (late fusion)	SAR + Optical	87.1%	75.3%
Multimodal (ours)	SAR + Optical + Atmosphere	91.7%	82.6%

Key Findings:

Flood Detection (Reactive):

- SAR critical for cloud-penetration, capturing flood extent even during storms
- Optical adds limited value during event (clouds) but helps baseline land cover
- Atmospheric data provides minimal direct detection benefit
- Multimodal fusion improves by 4-6 F1 points over SAR-only

Flood Prediction (Proactive):

- Atmospheric data most informative (precipitation forecasts)
- SAR/optical provide contextual land cover (urban vs. agricultural areas have different flood dynamics)
- Multimodal approach achieves 10+ point AUC improvement over any single modality
- 24-hour lead time: 82.6% AUC (usable for early warnings)
- 48-hour lead time: 76.3% AUC (degrades but still valuable)

Computational Cost:

- Multimodal model: 45M parameters (vs. 19M for Prithvi-256 image-only)
- Inference time: 12 seconds per tile (vs. 6 seconds image-only)
- Feasible for near-real-time edge deployment with modern accelerators

5.4 Future Directions: Integration with Physics Models

Current Limitations:

- AI models are data-driven, lack physical understanding
- Extrapolation beyond training distribution unreliable
- Cannot incorporate domain knowledge (fluid dynamics, thermodynamics)

Physics-Informed Neural Networks (PINNs):

- Hybrid models combining neural networks with differential equation solvers
- Loss functions include physics-based penalties (e.g., conservation laws)
- Example: Flood routing models constrained by shallow water equations

Potential Architecture:

1. Neural emulator: Fast surrogate for expensive physics simulation
2. Residual correction: NN learns deviation of real observations from physics model
3. Uncertainty quantification: Estimate confidence in predictions based on physics-data agreement

Advantages:

- Improved extrapolation (physics guides beyond training data)
- Reduced data requirements (physics provides inductive bias)
- Interpretability (model respects known physical laws)

Challenges:

- Computational cost (solving PDEs on-device is demanding)
- Model complexity (tuning hybrid systems requires both ML and domain expertise)
- Validation (ensuring NN doesn't exploit loopholes in approximate physics)

6. Barriers to Deployment and Recommended Strategies

6.1 Technical Barriers

Barrier	Impact	Proposed Mitigation
Data scarcity	Limits training of specialized models	Synthetic data generation, transfer learning from foundation models
Limited onboard compute	Restricts model complexity	Model compression (quantization, pruning), efficient architectures (MobileNets, EfficientNets)
Radiation-induced faults	Degrades inference accuracy	Hardware ECC, software checksums, ensemble redundancy

Communication bandwidth	Limits model updates	Differential uploads (only changed parameters), compressed task heads
Power constraints	Reduces inference throughput	Dynamic voltage scaling, scheduled inference during eclipse

6.2 Data and Benchmarking Needs

Current Gaps:

- Lack of standardized benchmarks: Different papers use incompatible datasets, metrics
- Limited geographic diversity: Most datasets focus on North America/Europe
- Temporal sparsity: Few datasets with dense time series for change detection
- Modality gaps: SAR, hyperspectral datasets lag behind optical

Recommended Actions:

1. Community Benchmark Initiatives:
    - Standardized train/val/test splits for key datasets (DOTA, DIOR, SIOR, etc.)
    - Common evaluation protocols and metrics
    - Regular leaderboard challenges (similar to ImageNet for natural images)
2. Data Collection Campaigns:
    - Partner with space agencies for coordinated data releases
    - Focus on underrepresented regions (Africa, Southeast Asia, South America)
    - Time-series datasets for disaster monitoring (before/during/after events)
3. Annotation Efficiency:
    - Invest in semi-supervised and self-supervised methods reducing labeling needs
    - Crowdsourcing platforms for large-scale annotation (quality control critical)
    - Active learning: Prioritize labeling images that most improve model performance

6.3 Model Sharing and Reproducibility

Challenges:

- Research code often incomplete, undocumented, or incompatible with deployment environments
- Trained models rarely released due to concerns about commercial sensitivity
- Computational requirements for training foundation models (millions of USD) prohibitive for most research groups

Recommendations:

1. Open Model Zoos:

    - Centralized repository of pre-trained geospatial AI models (analogous to Hugging Face for NLP)
    - Include model cards: Dataset, training procedure, performance metrics, known limitations
    - Versioning and provenance tracking
2. Deployment-Ready Formats:
    - Provide models in multiple formats: PyTorch, ONNX, OpenVINO, TensorFlow Lite
    - Include optimization scripts (quantization, pruning)
    - Document hardware requirements and expected performance
3. Reproducibility Standards:
    - Journals/conferences require code and model release for acceptance
    - Automated validation: Re-run training, verify reported metrics
    - Containerized environments (Docker) ensuring consistent execution

6.4 Collaboration Between Academia, Industry, and Space Agencies

Current Landscape:

- Academia: Cutting-edge algorithms, limited access to high-quality data and flight opportunities
- Industry: Operational expertise, proprietary datasets, risk-averse deployment
- Space Agencies: Mission platforms, long-term funding, coordination challenges

Recommended Collaboration Models:

Public-Private Partnerships:

- Co-funded research programs (e.g., ESA's ϕ-lab, SmartSat CRC model)
- Academia develops novel algorithms, industry/agency provides data and deployment pathways
- IP agreements clarified upfront

Hosted Payload Programs:

- Space agencies provide flight slots for academic/commercial AI experiments
- Standardized interfaces reduce integration costs
- Examples: ISS IMAGIN-e, Kanyini mission

Data Sharing Agreements:

- Tiered access: Open data for research, commercial licensing for operational use
- Embargo periods (e.g., 6-12 months) before public release
- Anonymization/redaction for sensitive regions

Joint Benchmarking Campaigns:

- Co-design evaluation protocols meeting both research and operational needs

- Regular competitions with real mission scenarios
- Prizes/contracts for top performers (incentivizes industry participation)

6.5 Standardization and Interoperability

Problem:

- Fragmented tooling: Different frameworks, formats, and interfaces across organizations
- Vendor lock-in: Proprietary hardware/software ecosystems
- Deployment friction: Models trained in one environment may not run in another

Proposed Standards:

Data Formats:

- Adopt open standards (e.g., Cloud-Optimized GeoTIFF, STAC metadata)
- Common coordinate systems and projections
- Metadata schemas for multimodal data (capture conditions, sensor specs)

Model Formats:

- ONNX for cross-framework compatibility
- Standard quantization formats (e.g., INT8, FP16)
- Defined input/output interfaces (image dimensions, channel ordering, pre/post-processing)

APIs and Protocols:

- RESTful APIs for on-orbit inference services
- gRPC for low-latency ground-to-space communication
- MQTT for telemetry and status reporting

Benefits:

- Accelerated deployment: Pre-validated models run on diverse platforms
- Reduced duplication: Shared tooling and infrastructure
- Lower barriers to entry: Smaller organizations can participate

7. Conclusions and Future Outlook

7.1 Key Takeaways

This workshop synthesized cutting-edge research addressing three critical challenges in space-based AI:

1. Data Scarcity: Demonstrated viable pathways for training robust models with limited labeled data through synthetic data generation (physics simulation,

generative AI) and foundation models enabling transfer learning. Synthetic SAR generation achieved 85%+ classification transfer accuracy, while foundation models reduced labeling needs by 50-75%.

2. AI Generalization: Showcased zero-shot learning techniques allowing models to recognize novel object categories and adapt to new environments without retraining. Context-aware architectures specifically designed for remote sensing improved detection mAP by 5-8 points over general-purpose vision models.
3. Edge Deployment: Validated compact geospatial foundation models on flight-representative hardware, achieving 5-6 second inference latency, 5-7W power consumption, and maintaining 92-98% of full-scale model accuracy. Successful deployment on ISS (IMAGIN-e) and Kanyini satellite demonstrates operational readiness.

7.2 Emerging Trends

Multimodal Foundation Models: The next generation of geospatial AI will integrate diverse data sources (SAR, optical, hyperspectral, atmospheric, sensor networks) within unified architectures. Self-supervised pre-training on massive multimodal datasets will enable:

- Better generalization across modalities (transfer learning from data-rich to data-scarce modalities)
- Improved robustness (degraded or missing modalities handled gracefully)
- Enhanced prediction (complementary information improves forecasting)

Early results demonstrate 10+ point accuracy improvements for disaster monitoring tasks when combining satellite imagery with atmospheric data.

Physics-Informed AI: Hybrid models incorporating domain knowledge (electromagnetic scattering theory, atmospheric dynamics, fluid mechanics) with data-driven learning will:

- Improve extrapolation beyond training distributions
- Reduce data requirements through physics-based inductive biases
- Enhance interpretability and trustworthiness

Challenges include computational cost (solving PDEs onboard) and complexity (requiring both ML and domain expertise).

Adaptive and Continual Learning: Foundation models enabling on-orbit learning will allow satellites to:

- Fine-tune models using in-situ data (adapt to specific geographic regions or seasons)
- Learn from user feedback (ground operators correct errors, model incorporates corrections)
- Handle distribution shift (Earth's surface evolves, models must track changes)

Technical challenges include limited onboard compute for training and catastrophic forgetting (new knowledge overwriting old).

7.3 Recommended Research Priorities

1. Efficient Model Architectures:
    - Develop transformer variants optimized for remote sensing (accounting for spatial invariance, multi-scale objects)
    - Explore mixture-of-experts architectures (activate relevant subnetworks based on scene content)
    - Hardware-aware neural architecture search (co-design models and accelerators)
2. Self-Supervised Learning:
    - Leverage temporal consistency (same location at different times should have consistent semantics)
    - Exploit multi-view geometry (different satellite viewing angles provide complementary information)
    - Cross-modal self-supervision (SAR and optical of same scene should have aligned features)
3. Trustworthy AI:
    - Uncertainty quantification (models must indicate confidence, especially for safety-critical applications)
    - Explainability (why did model make particular detection/classification?)
    - Robustness certification (formal guarantees of performance under adversarial conditions or bit flips)
4. Human-AI Collaboration:
    - Interactive labeling tools (AI suggests annotations, human refines)
    - Active learning (model requests labels for most informative examples)
    - Human-in-the-loop deployment (ground operators validate critical inferences)
5. End-to-End Mission Optimization:
    - Joint optimization of sensor tasking, onboard processing, and downlink scheduling
    - Reinforcement learning for autonomous satellite operations
    - Multi-satellite coordination (constellations sharing information, distributing computation)

7.4 Vision for 2030

Technical Capabilities:

- Orbital AI Processing: 90%+ of satellite imagery processed onboard, only high-value analytics downlinked
- Foundation Models as Standard: Every satellite deploys general-purpose encoder, mission-specific heads uploaded as needed
- Real-Time Disaster Response: Automated detection and alerting within 30 minutes of image capture

- Autonomous Constellations: Swarms of satellites coordinate observations, distribute computation, adapt to dynamic priorities

Ecosystem Maturity:

- Open Model Repositories: Comprehensive libraries of pre-trained geospatial AI models, freely available
- Standardized Interfaces: Plug-and-play model deployment across diverse satellite platforms
- Thriving Commercial Sector: Dozens of companies offering AI-as-a-service for satellite operators
- Global Accessibility: Developing nations benefit from advanced AI capabilities without building infrastructure

Societal Impact:

- Climate Monitoring: Continuous tracking of deforestation, ice melt, ocean health with meter-scale resolution
- Disaster Mitigation: Early warnings save thousands of lives annually, recovery efforts optimized through AI-guided damage assessment
- Sustainable Development: AI-driven insights inform policy, optimize resource management, monitor SDG progress
- Scientific Discovery: Automated analysis of vast satellite archives reveals previously hidden patterns and phenomena

7.5 Call to Action

The convergence of AI and space technology offers unprecedented opportunities to address global challenges. Realizing this potential requires:

- Sustained Investment: Multi-year funding for foundational research and demonstration missions
- Open Collaboration: Breaking down silos between academia, industry, and government
- Ethical Governance: Proactive development of guidelines for responsible AI deployment in space
- Capacity Building: Training next generation of researchers at the intersection of AI and space science
- Public Engagement: Communicating the value and possibilities of space-based AI to broader society

The workshop participants commit to advancing these priorities through continued research, open publication of models and datasets, and collaboration with the global community. The future of Earth observation is intelligent, autonomous, and accessible—the foundation has been laid, and the journey forward is underway.

Acknowledgments

This white paper synthesizes presentations and discussions from the AI for Space and Remote Sensing workshop. The authors acknowledge:

- University of Adelaide AI for Space Group: Andrew Du, Roberto del Prete (ESA ϕ-lab), Nick Manser, Fabrice Marre, Andrew Barton, Carl Seubert, Gabriele Meoni (ESA ϕ-lab), and Prof. Tat-Jun Chin
- Nanyang Technological University CISS-ROSE AI Centre: Assoc. Prof. Bihan Wen and research team
- Funding Support: SmartSat CRC, Australian Institute for Machine Learning, ESA ϕ-lab, UoA-UNSW Defence Trailblazer, NTU Satellite Research Centre
- Mission Partners: Kanyini mission consortium, Thales Alenia Space, Microsoft Azure Orbital

References

Key publications from the workshop presentations:

1. Qing et al., "EO to SAR Image Generation," Remote Sensing, 2023
2. Huang, Wen et al., "Zero-Shot Remote Sensing Instance Segmentation," AAAI, 2024
3. Huang, Wen et al., "Region-Aware Semantic Context Integration for Zero-Shot Detection," ICCV, 2025
4. Du et al., "Compact Geospatial Foundation Models via Knowledge Distillation," (In preparation)

Full references available from presenting authors.