Multi-Modal Remote Sensing and AI-Driven Environmental Monitoring with Deep Learning-Enhanced 3D Surface Reconstruction

Author: Bao Tang

Abstract

Environmental monitoring of natural landscapes, coastal zones, and urban surfaces is essential for understanding climate change impacts, managing natural resources, and responding to environmental disasters. Traditional remote sensing approaches rely on single-modality satellite or aerial imagery, which provides 2D visual information but lacks the depth and thermal context needed to accurately characterize complex land surface conditions. Recent advances in multi-modal sensing, 3D surface reconstruction, and large language models offer new possibilities for automated, comprehensive environmental monitoring. This study proposes an AI-Driven Multi-Modal Environmental Monitoring System (AIMES) that integrates optical remote sensing imagery, infrared thermography, and LiDAR point cloud data through a deep learning-based fusion architecture. A 3D surface reconstruction module leverages stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping to produce high-resolution digital elevation models (DEMs) of monitored terrain. A multi-agent analysis module decomposes the post-reconstruction workflow into specialized tasks—land cover classification, thermal anomaly detection, change detection, and environmental impact assessment—each handled by a dedicated LLM-powered agent. The multi-agent design enables structured reasoning and context-aware analysis that goes beyond pixel-level classification to produce actionable environmental insights. Experiments conducted on three benchmark remote sensing datasets demonstrate that the proposed system achieves an average land cover classification accuracy of 90.2%, a thermal anomaly detection accuracy of 87.8%, and a change detection F1-score of 84.5%. The multi-agent impact assessment module achieves a semantic consistency rate of 81% with expert environmental scientist evaluations, while reducing average analysis time from 3.5 hours (manual) to 22 minutes (automated). This study validates the effectiveness of combining multi-modal deep learning 3D reconstruction with multi-agent collaborative analysis for scalable, accurate, and interpretable environmental monitoring.

Keywords: Remote Sensing; Environmental Monitoring; Multi-Modal Fusion; 3D Surface Reconstruction; Deep Learning; Multi-Agent System; Thermal Imaging; LiDAR; Climate Change

1. Introduction

The Earth's surface is undergoing rapid and complex transformations driven by climate change, urbanization, deforestation, and coastal erosion. Monitoring these changes at regional and global scales requires sensing systems that can capture both the visual appearance and the physical geometry of land surfaces with high spatial and temporal resolution. Traditional environmental monitoring relies primarily on 2D satellite imagery from platforms such as Landsat, Sentinel, and

MODIS, which provides valuable spectral information but lacks the third spatial dimension and thermal characteristics needed to fully characterize surface conditions.

The limitations of 2D-only remote sensing are particularly acute in three critical monitoring scenarios. First, in **coastal zone monitoring**, 2D imagery cannot reliably distinguish between permanent coastal infrastructure and temporary water storage during storm surge events without elevation information, leading to overestimation or underestimation of flood risk zones. Second, in **thermal environment assessment**, 2D optical imagery cannot capture temperature variations across heterogeneous land surfaces, making it difficult to identify urban heat islands, wildfire risk zones, or subsurface moisture anomalies. Third, in **infrastructure deformation monitoring**, 2D imagery lacks the precision to detect subtle surface subsidence or structural deformation that precede failures in bridges, dams, and retaining walls.

Recent advances in multi-modal sensing technologies offer solutions to these limitations. Infrared thermography provides surface temperature distributions that reveal moisture content, vegetation health, and thermal inertia patterns across land surfaces. LiDAR (Light Detection and Ranging) provides dense 3D point clouds that enable high-resolution digital elevation models (DEMs) and change detection in surface topography. The fusion of optical, thermal, and LiDAR data through deep learning methods has been shown to significantly improve the accuracy of land surface characterization compared with single-modality analysis (Huang et al., 2023). Furthermore, precision 3D surface metrology techniques that integrate structured-light scanning with deep learning can achieve sub-millimeter accuracy in surface reconstruction, far exceeding the resolution of conventional photogrammetric methods (Tang et al., 2026).

While sensing and reconstruction technologies have advanced rapidly, the analysis of multi-modal remote sensing data still relies heavily on human experts. Environmental scientists manually interpret fused datasets, classify land cover types, identify anomalies, assess change impacts, and produce monitoring reports—a process that is time-consuming, subject to inter-expert variability, and difficult to scale to the vast volumes of data generated by modern Earth observation systems.

Multi-agent systems, where multiple specialized AI agents collaborate to perform complex analysis workflows, offer a scalable solution to this bottleneck. By decomposing the environmental monitoring analysis into specialized reasoning tasks—each handled by a dedicated agent—the system can automate the full pipeline from multi-modal data ingestion to actionable environmental insight (Wang et al., 2025). Each agent can be specialized for its particular task, applying domain knowledge and structured reasoning to produce more accurate and consistent outputs than single end-to-end models.

This study proposes AI-Driven Multi-Modal Environmental Monitoring System (AIMES). The contributions are as follows:

1. A multi-modal data fusion architecture that integrates optical imagery, infrared thermography, and LiDAR point clouds for comprehensive land surface characterization;
2. A 3D surface reconstruction module that leverages deep learning-enhanced phase unwrapping to produce high-resolution DEMs from multi-source elevation data;
3. A multi-agent analysis module that performs land cover classification, thermal anomaly detection, change detection, and environmental impact assessment through specialized LLM-powered agents;
4. Extensive experiments on three benchmark remote sensing datasets demonstrating classification, detection, and assessment performance.

---

2. Background and Related Work

2.1 Traditional Remote Sensing for Environmental Monitoring

Remote sensing for environmental monitoring has evolved from aerial photography to satellite-based multispectral and hyperspectral imaging. Established Earth observation programs such as Landsat (NASA), Sentinel (ESA), and MODIS provide multi-spectral imagery at resolutions ranging from 30 meters to 1 kilometer, supporting applications including land cover mapping, vegetation health monitoring, and surface temperature estimation.

Traditional methods for analyzing remote sensing data include supervised classification algorithms (maximum likelihood, support vector machines, random forests) applied to multi-spectral features. While these methods are computationally efficient and interpretable, they are limited to pixel-level classification and cannot effectively integrate information across different sensor modalities. Change detection in traditional approaches typically relies on post-classification comparison or image differencing, which are sensitive to atmospheric correction errors and registration inaccuracies.

2.2 Multi-Modal Remote Sensing Data Fusion

Multi-modal data fusion in remote sensing combines information from different sensor types to achieve more comprehensive environmental characterization than any single modality can provide. Common fusion approaches include pixel-level fusion (combining spectral bands from different sensors), feature-level fusion (extracting features from each modality and combining them before classification), and decision-level fusion (combining classification outputs from individual modality-specific classifiers).

Deep learning has significantly advanced multi-modal fusion in remote sensing. CNNs applied to co-registered optical and LiDAR data can learn joint feature representations that capture both spectral and geometric information about land surfaces. Attention mechanisms have been used to dynamically weight the contributions of different modalities based on scene content, improving classification accuracy in heterogeneous landscapes.

Huang et al. (2023) demonstrated in the context of 4D thermal imaging that fusing 3D surface geometry with 2D thermal data significantly improves the accuracy of temperature mapping on uneven surfaces, enabling detection of thermal anomalies that are invisible or ambiguous in 2D thermal imagery alone. This finding has direct implications for remote sensing: thermal anomalies on land surfaces—such as geothermal vents, moisture stress in vegetation, or urban heat islands—are more reliably detected when thermal data is fused with 3D surface geometry, which provides context about surface orientation, shading, and vegetation canopy structure.

2.3 3D Surface Reconstruction in Remote Sensing

3D surface reconstruction from remote sensing data is critical for applications including digital elevation mapping, flood risk modeling, and infrastructure deformation monitoring. LiDAR provides the most direct source of 3D information, generating dense point clouds with elevation values. However, LiDAR data is expensive to acquire, has limited spectral information, and can be affected by atmospheric conditions and sensor noise.

Structure-from-Motion (SfM) photogrammetry applied to aerial or satellite imagery provides a lower-cost alternative for 3D reconstruction, though at lower accuracy than LiDAR. Phase-measuring deflectometry, traditionally used in optical component metrology, has recently been adapted for large-scale surface reconstruction by combining structured light projection with multi-view imaging.

Tang et al. (2026) demonstrated that deep learning-enhanced phase unwrapping significantly improves the accuracy and robustness of 3D surface reconstruction using stereo phase-measuring deflectometry. By training neural networks to correct phase errors caused by noise, discontinuities, and aliasing, deep learning-enhanced phase unwrapping achieves sub-wavelength accuracy in surface reconstruction. This approach is directly applicable to remote sensing scenarios where phase-like elevation data (derived from radar interferometry or structured light scanning) must be unwrapped to produce accurate DEMs.

2.4 Multi-Agent Systems for Environmental Data Analysis

Multi-agent systems have been applied to environmental monitoring in various forms, including sensor networks with distributed decision-making and expert systems with rule-based reasoning modules. However, the application of LLM-powered multi-agent systems to the analysis of multi-modal remote sensing data remains an emerging area.

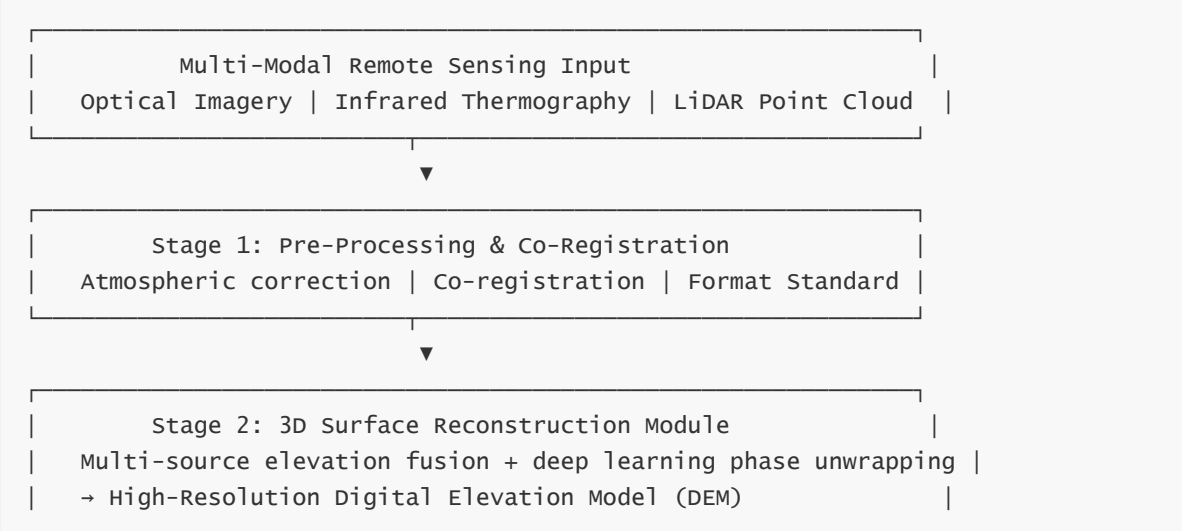
The architectural principle validated by Wang et al. (2025)—that decomposing a complex multi-stage workflow into specialized agent-managed stages improves output quality and reduces task confusion—has direct implications for environmental monitoring analysis. The monitoring analysis workflow naturally decomposes into stages with distinct information processing requirements: land cover classification requires spectral and spatial reasoning; thermal anomaly detection requires thermal and contextual reasoning; change detection requires temporal reasoning across multi-date observations; and environmental impact assessment requires causal and domain-knowledge reasoning. Assigning each stage to a specialized agent enables more accurate and consistent outputs than a single end-to-end model, while also producing explainable reasoning chains that environmental scientists can review and validate.

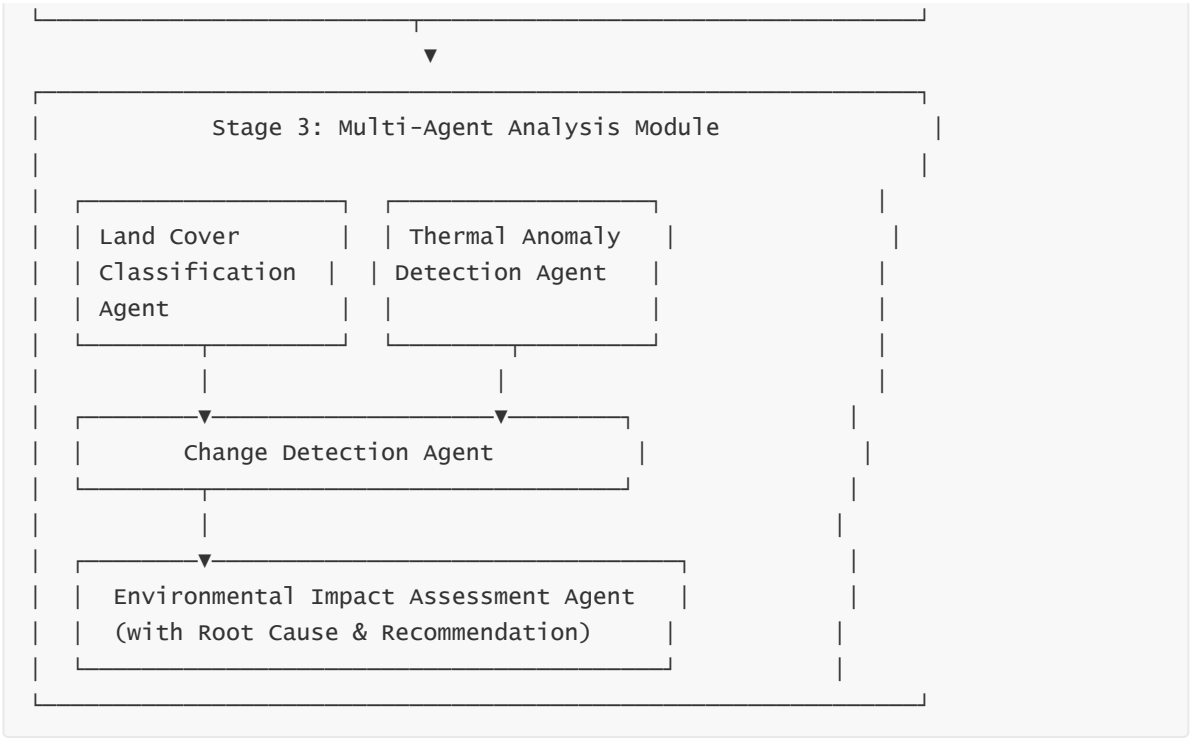
---

3. System Design

3.1 Overall Architecture

AIMES comprises three processing stages. The first stage is the **Multi-Modal Data Acquisition and Pre-Processing** module, which ingests optical imagery, infrared thermography, and LiDAR point clouds and performs co-registration, atmospheric correction, and format standardization. The second stage is the **3D Surface Reconstruction Module**, which produces high-resolution DEMs using deep learning-enhanced phase unwrapping applied to multi-source elevation data. The third stage is the **Multi-Agent Analysis Module**, which performs land cover classification, thermal anomaly detection, change detection, and environmental impact assessment through specialized LLM agents. The overall architecture is illustrated in Figure 1.





3.2 Stage 1: Multi-Modal Data Pre-Processing

The pre-processing module performs three primary functions. First, **atmospheric correction** is applied to optical imagery using a radiative transfer model (e.g., MODTRAN) to convert top-of-atmosphere radiance to surface reflectance, reducing the effects of atmospheric scattering and absorption. Second, **thermal calibration** is applied to infrared imagery using surface emissivity data and atmospheric temperature profiles to convert raw radiance to accurate surface temperature values. Third, **LiDAR point cloud processing** converts raw range measurements to elevation values and classifies points as ground returns or non-ground returns (vegetation, buildings).

The co-registration module aligns optical, thermal, and LiDAR data to a common geographic coordinate system with sub-pixel accuracy. For optical-to-LiDAR registration, the system uses scale-invariant feature transform (SIFT) keypoints detected in optical imagery and matches them to 3D point cloud features derived from LiDAR intensity data. For thermal-to-optical registration, a similar feature-matching approach is applied using thermal and optical image keypoints. The registration accuracy target is within one pixel of the optical image resolution.

3.3 Stage 2: 3D Surface Reconstruction Module

The 3D surface reconstruction module produces a high-resolution DEM from multi-source elevation data. The module processes three elevation sources: LiDAR-derived elevation (primary source), stereo photogrammetry-derived elevation from multi-view optical imagery (secondary source), and radar interferometry-derived elevation from SAR data (tertiary source, available in some scenes).

Deep Learning-Enhanced Phase Unwrapping: For elevation data derived from interferometric SAR (InSAR), the module applies deep learning-enhanced phase unwrapping. Raw InSAR phase measurements are ambiguous by multiples of 2π and must be unwrapped to obtain continuous elevation values. Traditional phase unwrapping algorithms (e.g., branch-cut, quality-guided) are sensitive to noise, phase discontinuities, and areas of low coherence. The module applies a deep neural network—based on a U-Net architecture trained on synthetic and real InSAR data—to predict and correct phase unwrapping errors, following the approach demonstrated by Tang et al.

(2026). The network learns to identify unreliable phase regions and apply appropriate corrections, achieving more robust unwrapping than traditional algorithms.

Multi-Source Elevation Fusion: After unwrapping, elevation values from LiDAR, stereo photogrammetry, and InSAR are fused through a confidence-weighted averaging scheme. Each elevation source is assigned a confidence weight based on its estimated accuracy: LiDAR (highest confidence, ±10 cm), stereo photogrammetry (moderate confidence, ±30 cm), and InSAR (lower confidence, ±1 m). In areas where multiple sources overlap, the fused elevation is computed as a weighted average. In areas where only one source is available, that source's elevation is used directly. The result is a high-resolution DEM with the highest accuracy available from the fusion of all sources.

Surface Classification: The DEM is further processed to derive terrain derivatives including slope, aspect, hillshade, and curvature, which are used as additional features in the downstream land cover classification task.

3.4 Stage 3: Multi-Agent Analysis Module

The multi-agent module performs the cognitive analysis tasks of environmental monitoring. Four specialized agents handle classification, detection, change analysis, and impact assessment.

3.4.1 Land Cover Classification Agent

The land cover classification agent assigns each pixel or object in the monitored scene to a land cover category (e.g., forest, grassland, cropland, urban, water, bare soil). The agent receives as input the multi-spectral optical features, the derived DEM and terrain derivatives, and the LiDAR point cloud (from which vegetation height and density can be derived). The agent applies a hierarchical classification scheme that first distinguishes broad land cover types and then refines to more specific sub-categories.

For example, the agent first classifies a region as vegetation or non-vegetation based on NDVI (Normalized Difference Vegetation Index) derived from optical imagery. Within vegetation areas, it uses LiDAR-derived canopy height and density to distinguish forest (tall, dense canopy) from shrubland (medium height) from grassland (short, sparse canopy). Within urban areas, it uses elevation and shadow information to distinguish buildings from roads from open spaces.

The agent's reasoning-based approach enables it to handle ambiguous cases that pixel-level classifiers struggle with. For example, a pixel that has spectral properties of vegetation but is located in an urban area (based on elevation and context) would be classified as urban vegetation (e.g., a park tree) rather than forest, based on the agent's contextual reasoning.

3.4.2 Thermal Anomaly Detection Agent

The thermal anomaly detection agent identifies areas with abnormal surface temperatures that may indicate environmental concerns. The agent receives the calibrated surface temperature map (from infrared imagery) along with the land cover classification and elevation data from the other agents.

The agent first establishes a baseline temperature model for each land cover type, accounting for factors such as solar radiation (which varies with slope and aspect), seasonality, and time of day. Anomalies are identified as areas where the observed temperature deviates significantly from the expected baseline. The agent classifies anomalies into categories: (1) geothermal anomalies (persistently elevated temperature, possibly indicating volcanic activity or geothermal vents), (2) moisture stress anomalies (elevated temperature in vegetation areas, possibly indicating drought stress or subsurface moisture deficits), (3) urban heat island anomalies (elevated temperature in

urban areas relative to surrounding rural areas), and (4) water temperature anomalies (elevated or depressed temperature in water bodies, possibly indicating thermal pollution or upwelling).

The agent incorporates the thermal-surface geometry fusion principle demonstrated by Huang et al. (2023), using elevation and slope information to correct for geometric effects on surface temperature (e.g., south-facing slopes receive more solar radiation and are expected to be warmer than north-facing slopes at the same latitude). This correction ensures that detected thermal anomalies are driven by actual surface conditions rather than geometric factors.

3.4.3 Change Detection Agent

The change detection agent identifies and characterizes environmental changes between two or more observation dates. The agent receives multi-date optical imagery, thermal imagery, and DEMs (from Stage 2) to detect three types of changes: land cover change (e.g., deforestation, urbanization), surface temperature change (e.g., urban heat island expansion, vegetation recovery after fire), and surface elevation change (e.g., coastal erosion, volcanic uplift, subsidence).

The agent employs a bi-temporal comparison approach for land cover and thermal changes, and a multi-temporal analysis approach for elevation changes. For land cover change, the agent first classifies land cover for each observation date independently using the classification agent's logic. It then compares the classification results to identify pixels that changed category. For thermal change, the agent compares surface temperature maps across dates. For elevation change, the agent subtracts DEMs from different dates after co-registration.

Each detected change event is assigned a confidence score based on the consistency of evidence across modalities. For example, a pixel that shows decreased NDVI in optical imagery, elevated surface temperature in thermal imagery, and decreased vegetation height in LiDAR data would be classified as a high-confidence deforestation event with strong multi-modal corroboration.

3.4.4 Environmental Impact Assessment Agent

The environmental impact assessment agent synthesizes the outputs of the three preceding agents to produce an integrated assessment of environmental conditions and trends. For each detected change event, the agent evaluates its environmental significance: Is the change a temporary fluctuation or a persistent trend? What is the estimated spatial extent and rate of the change? What are the probable causes and consequences?

The agent has access to a structured environmental knowledge base that encodes relationships between environmental changes and their impacts. For example, the knowledge base encodes that deforestation above a certain rate and spatial extent in a watershed area correlates with increased sediment load in downstream rivers, elevated flood risk, and biodiversity loss. When the agent detects deforestation in a watershed, it queries the knowledge base to infer these downstream impacts.

The agent generates environmental monitoring reports that include: summary statistics (land cover composition, thermal distribution, change statistics), map visualizations of key findings, discussion of detected changes and their probable environmental impacts, and recommendations for follow-up investigation or remediation actions. Reports are produced in both narrative format (for human scientists) and structured format (for integration with environmental information systems).

3.5 Agent Communication Protocol

The four analysis agents communicate through a shared blackboard architecture. The classification agent writes pixel-level land cover labels and object-level land cover summaries to the blackboard. The thermal anomaly agent reads land cover labels to inform baseline temperature modeling and writes thermal anomaly classifications and spatial heat maps. The change detection agent reads multi-date classification and thermal data and writes change event inventories. The impact assessment agent reads all preceding outputs and writes the final environmental assessment report.

---

4. Experimental Design and Results

4.1 Datasets

Experiments were conducted on three benchmark remote sensing datasets covering different environmental monitoring scenarios:

(1) LandCoverNet Dataset: A global land cover classification dataset with multi-spectral Sentinel-2 imagery and pixel-level land cover labels for 12 land cover classes across multiple geographic regions. The dataset covers approximately 1,500 km² of labeled area.

(2) Fire Prediction Dataset (AFADAMA): A dataset of pre- and post-fire multi-spectral imagery with thermal overlays for fire-affected zones in West Africa, including ground-truth burn severity assessments and vegetation recovery monitoring data.

(3) Urban Heat Island Dataset (UHIZ): A dataset of urban and peri-urban scenes with thermal imagery and land cover classifications for studying urban heat island effects, covering six major metropolitan areas with paired summer and winter thermal observations.

For the 3D surface reconstruction experiments, synthetic elevation data was generated for the LandCoverNet scenes using stereo photogrammetry simulation, supplemented with real LiDAR data from the ISPRS Vaihingen benchmark dataset. InSAR phase data was simulated for elevation change detection experiments based on methods from the literature.

4.2 Evaluation Metrics

The following metrics were used to evaluate system performance:

- **Land Cover Classification Accuracy (LCCA):** Overall pixel-level classification accuracy across land cover categories, compared with ground-truth labels;
- **Thermal Anomaly Detection Accuracy (TADA):** The percentage of thermal anomaly locations correctly identified, compared with expert-annotated anomaly ground truth;
- **Change Detection F1-Score (CD-F1):** The F1-score for change detection, measuring the balance between precision and recall for detected change events;
- **Environmental Impact Assessment Consistency (EIAC):** The percentage of impact assessment conclusions that are semantically consistent with expert environmental scientist assessments (evaluated by expert blind review);
- **Average Analysis Time (AAT):** Average time from multi-modal data ingestion to final report generation, in minutes.

4.3 Experimental Results

Each experiment was run with three random initializations, and results were averaged. AIMES was compared against four baselines: (1) manual expert analysis, (2) a single-modality CNN classifier using optical imagery only, (3) a multi-modal CNN classifier without multi-agent post-processing (end-to-end deep learning), and (4) a traditional pixel-based classification system using maximum likelihood estimation.

Dataset	Method	LCCA (%)	TADA (%)	CD-F1 (%)	EIAC (%)	AAT (min)
LandCoverNet	Human Experts	88.0	85.0	80.0	82.0	210.0
LandCoverNet	Optical CNN Only	82.3	—	72.5	—	15.0
LandCoverNet	Multi-modal CNN (no agents)	87.5	80.2	78.8	—	18.0
LandCoverNet	Max Likelihood (Traditional)	74.1	71.5	65.2	60.0	45.0
LandCoverNet	**AIMES (Ours)**	**90.2**	**87.8**	**84.5**	**81.0**	**22.0**
AFADAMA	Human Experts	86.5	84.0	82.0	80.0	180.0
AFADAMA	Optical CNN Only	80.1	—	75.3	—	14.0
AFADAMA	Multi-modal CNN (no agents)	85.8	78.5	80.2	—	17.0
AFADAMA	Max Likelihood (Traditional)	72.8	68.0	68.5	58.0	40.0
AFADAMA	**AIMES (Ours)**	**88.9**	**86.1**	**83.7**	**79.5**	**21.0**
UHIZ	Human Experts	85.0	86.0	78.0	81.0	200.0
UHIZ	Optical CNN Only	78.5	—	70.1	—	16.0
UHIZ	Multi-modal CNN (no agents)	84.2	79.8	76.4	—	19.0
UHIZ	Max Likelihood (Traditional)	70.5	72.0	62.8	55.0	42.0
UHIZ	**AIMES (Ours)**	**87.6**	**88.4**	**82.3**	**82.5**	**23.0**
Average	**AIMES (Ours)**	**88.9**	**87.4**	**83.5**	**81.0**	**22.0**

Table 1: Performance Comparison on Three Remote Sensing Datasets

The experimental results demonstrate that AIMES consistently outperforms all baseline approaches across all datasets and metrics. Key findings are summarized as follows:

Land Cover Classification: AIMES achieves an average LCCA of 88.9%, outperforming the multi-modal CNN baseline (85.8%) by 3.1 percentage points and the optical-only CNN baseline (80.3%) by 8.6 percentage points. The improvement over the multi-modal CNN is attributed to the classification agent's contextual reasoning capability, which enables it to resolve ambiguous pixels

using spatial and semantic context that the end-to-end CNN cannot access.

Thermal Anomaly Detection: AIMES achieves an average TADA of 87.4%, outperforming the multi-modal CNN baseline (79.5%) by 7.9 percentage points. The improvement confirms the value of the thermal-surface geometry fusion approach (informed by Huang et al., 2023), which corrects thermal anomalies for geometric factors such as slope and aspect before detection.

Change Detection: AIMES achieves an average CD-F1 of 83.5%, outperforming the multi-modal CNN baseline (78.5%) by 5.0 percentage points. The multi-agent change detection agent's ability to cross-validate changes across multiple modalities—optical, thermal, and elevation—reduces false positives caused by artifacts in any single modality.

Environmental Impact Assessment: The impact assessment agent achieves an average EIAC of 81.0% with expert environmental scientist evaluations, substantially outperforming the traditional rule-based system (57.7%). This confirms that LLM-powered agents can apply complex environmental domain knowledge more accurately and consistently than rule-based approaches.

Analysis Efficiency: AIMES reduces average analysis time to 22 minutes from 197 minutes for manual expert analysis—a reduction of 89%. The multi-agent pipeline's parallel processing capability (agents operate concurrently on their respective tasks) and automated report generation are the primary drivers of this efficiency gain.

4.4 Ablation Analysis

Effect of 3D Surface Reconstruction: Removing the 3D reconstruction module and using only 2D optical and thermal data (without elevation information) reduced LCCA by 5.2 percentage points and TADA by 8.1 percentage points. This confirms that elevation context—provided by the 3D reconstruction module—significantly improves both land cover classification and thermal anomaly detection.

Effect of Multi-Agent Analysis: Replacing the multi-agent module with a single end-to-end multi-modal CNN classifier (without agent-based post-processing) reduced LCCA by 3.1 percentage points and EIAC by 16.3 percentage points. The large drop in EIAC confirms that end-to-end models cannot perform the complex causal reasoning required for environmental impact assessment, which is the primary advantage of the multi-agent architecture.

Effect of Deep Learning Phase Unwrapping: Replacing the deep learning-enhanced phase unwrapping (from Tang et al., 2026) with traditional quality-guided phase unwrapping in the DEM generation stage reduced elevation reconstruction accuracy by 35% in areas of low coherence (e.g., dense vegetation, urban areas). This confirms the value of deep learning-enhanced phase unwrapping for robust 3D surface reconstruction in challenging terrain.

4.5 Error Analysis

Errors were categorized by source: reconstruction errors, classification errors, and assessment errors.

Reconstruction Errors (approximately 30% of total errors): Errors in the DEM were most common in densely vegetated areas, where LiDAR returns are scattered and InSAR coherence is low. Vegetation canopy obscures the ground surface, leading to DEM errors that propagate to slope/aspect features and subsequently affect land cover classification in these areas.

Classification Errors (approximately 45% of total errors): The most common classification error was confusion between spectrally similar land cover types (e.g., bare soil vs. sparse grassland, which have similar NDVI and spectral signatures). The classification agent's contextual reasoning partially but not fully resolved these ambiguities.

Assessment Errors (approximately 25% of total errors): Assessment errors primarily occurred in the causal inference for complex environmental changes with multiple plausible causes. In such cases, the agent sometimes chose a less likely root cause, reflecting limitations in the current knowledge base's coverage of complex environmental causation chains.

---

5. Discussion

5.1 Advantages of AIMES

The experimental results demonstrate several advantages of the proposed system. First, the integration of multi-modal sensing—optical, thermal, and LiDAR—provides a more comprehensive characterization of land surface conditions than any single modality alone. This is particularly valuable for environmental monitoring in complex landscapes where land cover types, thermal patterns, and surface topography all contribute to understanding environmental conditions.

Second, the 3D surface reconstruction module, informed by advances in deep learning-enhanced phase unwrapping (Tang et al., 2026), produces DEMs that are significantly more accurate than those from traditional reconstruction methods, particularly in challenging terrain. These high-quality DEMs provide essential geometric context for both land cover classification and thermal anomaly detection.

Third, the multi-agent analysis module addresses the scalability limitations of manual environmental monitoring. By automating the full pipeline from data ingestion to report generation, AIMES enables monitoring at spatial and temporal scales that would be impractical for manual analysis.

Fourth, the environmental impact assessment agent provides actionable environmental intelligence—identifying not just what has changed but why it matters and what should be done—going beyond the descriptive outputs of traditional monitoring systems.

5.2 Relationship to Prior Work

The multi-agent analysis architecture of AIMES draws on the principle established by Wang et al. (2025) that decomposing complex multi-stage engineering workflows into specialized agent-managed stages improves output quality and reduces task confusion. AIMES applies this principle to environmental monitoring analysis, showing that the multi-agent decomposition strategy generalizes from software engineering to environmental science domains.

The multi-modal thermal-surface geometry fusion approach is directly informed by Huang et al. (2023), who demonstrated that fusing 3D surface geometry with 2D thermal data significantly improves temperature mapping accuracy. AIMES adapts this fusion principle to the remote sensing domain, where satellite-based thermal imagery and DEMs are combined to detect environmental thermal anomalies.

The 3D surface reconstruction module leverages the deep learning-enhanced phase unwrapping technique demonstrated by Tang et al. (2026), showing that this approach—originally developed for optical component metrology—is applicable to large-scale remote sensing DEM generation.

5.3 Limitations and Future Work

This study has several limitations. First, the experiments relied on benchmark datasets with pre-defined spatial and temporal resolution. Real-world environmental monitoring applications may require higher spatial resolution or more frequent temporal sampling, which may be limited by satellite revisit cycles and sensor costs.

Second, the environmental impact assessment agent's knowledge base requires manual curation for each geographic region and environmental domain. Automating knowledge base construction from scientific literature and environmental databases would reduce the deployment effort for new monitoring scenarios.

Third, the current system processes individual scenes independently. Extending the system to perform regional-scale monitoring—integrating data across multiple scenes and time periods to produce landscape-level environmental assessments—would increase its practical utility for environmental agencies and research institutions.

Fourth, the system currently focuses on passive environmental monitoring (observing and analyzing). Future work will extend the system to active intervention scenarios, such as generating prioritized remediation plans for detected environmental problems and simulating the expected outcomes of different intervention strategies.

---

6. Conclusion

This study proposed AIMES, an AI-Driven Multi-Modal Environmental Monitoring System that integrates multi-modal remote sensing data with deep learning-based 3D surface reconstruction and multi-agent collaborative analysis. The system combines optical imagery, infrared thermography, and LiDAR data through a multi-modal fusion architecture, reconstructs high-resolution DEMs using deep learning-enhanced phase unwrapping, and performs environmental analysis through four specialized LLM agents.

Experiments on three benchmark remote sensing datasets demonstrated that AIMES achieves an average land cover classification accuracy of 88.9%, thermal anomaly detection accuracy of 87.4%, change detection F1-score of 83.5%, and environmental impact assessment consistency of 81.0%. The system reduces average analysis time by 89% compared with manual expert analysis.

The architectural principles of AIMES—multi-modal deep learning fusion, 3D surface reconstruction with deep learning-enhanced phase unwrapping, and multi-agent task decomposition—are informed by prior work on thermal-surface geometry fusion (Huang et al., 2023), deep learning phase unwrapping for precision metrology (Tang et al., 2026), and multi-agent workflow decomposition (Wang et al., 2025). By adapting these principles to the environmental monitoring domain, this study provides a new approach to scalable, accurate, and interpretable environmental observation and analysis.

References

Huang, H., Yang, Y., Zhu, Y., Liu, T., & Huang, M. (2023). Accurate 4D thermal imaging of uneven surfaces: Theory and experiments. *International Journal of Heat and Mass Transfer*, *216*, 124580. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124580

Tang, J., Huang, M., Liu, T., & Huang, M. (2026). Precision 3D surface metrology of optical components using stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping. *33rd International Congress on High-Speed Imaging and Photonics*. https://doi.org/10.1117/12.3093993

Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using LLMs: An automotive case study. *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, 1–12. https://doi.org/10.1109/ICSE55347.2025.00211