Vol. 06, No. SI*, pp. 123 –134 (2025)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Effective Image and Video Recognition Techniques in Environmental and Earth Monitoring Systems Using Remote-Sensed Intelligent Visual Analytics

Gnana Rubini R[1*], J. Raja[2], P. Kamaraja pandian[3], Vilas Namdeo Nitnaware[4], Sheemona Joseph[5], S. Ponni Alias Sathya[6]

[1]*Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore. gnanrajrad@gmail.com*
[2]*Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai-600062. drrajaj@veltech.edu.in*
[3]*Department of Computer Science and Engineering, Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, Hyderabad 501301. kamarajapandianp@gmail.com*
[4]*MAEER'S MIT Thane, Near Green Valley Studio, Mira Road, Kashi gaon, Mumbai, Maharashtra - 401107. vilasan30@yahoo.com*
[5]*Department of Cyber Security, Sri Krishna College of Engineering and Technology, Coimbatore. sheemonajoseph@gmail.com*
[6]*Department of Information Technology, Dr.Mahalingam College of Engineering and Technology, Pollachi. sathyaashok2007@gmail.com*

**Correspondence: gnanrajrad@gmail.com*

Abstract

Earth and environmental monitoring are very crucial to identify changes in climatic conditions, destruction of an ecosystem and calamities. The increased access to high-resolution satellite, aerial, and UAV imagery requires sophisticated intelligent visual analytics that can be used to derive actionable information on the basis of massive streams of remote-sensed data. The current image and video recognition methods are not always capable of attaining reliable performances in the presence of multimodal data heterogeneity, environmental dynamics, and interference of noise in remote-sensing images. These issues restrict the precision and flexibility of traditional deep learning-based monitoring systems to real-life applications. In this paper, we have suggested the Enhanced Visual Intelligence for Adaptive Recognition Network (EVIAR-Net). This deep learning model is a hybrid one that uses Graph-Convolutional Vision Transformers (GCVT) and Adaptive Multi-Source Fusion (AMSF). EVIAR-Net is able to store spatial correlations along with temporal dependencies using the graph-based spatial reasoning and transformer-based temporal encoding. AMSF actively combines multispectral, hyperspectral and video modalities to provide resistance to illumination, motion, and atmospheric perturbations. Performance assessments of various Earth observation datasets indicate an improvement in recognition accuracy of 21 percent, inference speed of 30 percent, and generalisation to unknown environments are better than CNN, ViT, and LSTM-based models. The suggested EVIAR-Net concept exhibits a smart, adaptable, and energy-saving strategy towards the next-generation environmental monitoring and predictive analytics.

Keywords: *Remote Sensing, Visual Analytics, Graph-Convolutional Transformer, Environmental Monitoring, Multimodal Fusion, Deep Learning.*

Received: October 02nd, 2025 / Revised:Decemeber 18th, 2025 / Accepted: December 27th, 2025 / Online: December 31st, 2025

I. INTRODUCTION

Visual intelligence is vital for environmental and earth monitoring practices, allowing computers to interpret satellite, aerial, and drone imagery to assist in ecological feedback and decision making [1]. Applying artificial intelligence (AI) and remote sensing enhances visual pattern recognition and understanding [2]. Image and video recognition methods respond to classifications in land use, damage assessment from disasters, and ecosystem monitoring [3]. Enhanced, high-resolution imaging and continuous-data input from a multitude of sensors deliver real-time environmental information about climate change, landscape changes, and other dynamics [4].

Intelligent visual analytics becomes a vehicle for converting massive amounts of visual data into meaningful environmental indicators [5]. Advanced transition methods of mapping

***SI*:** **Special issue - Remote Sensing based Intelligent Visual Analytics for Real-time Environmental and Earth Monitoring Systems**

doi:10.38094/jastt605659

vegetation loss, land-use change, soil degradation, water pollution, etc., cost-effectively maintain sustainable systems [6]. Visual intelligence also supports policymakers and researchers to practice predictive analysis and spatial-scale sustainable planning [7]. AI visual intelligence, combining deep learning and computer vision with remote sensing, can provide high accuracy and automation of computation and data interpretation [8]. Visual intelligence underpins the next generations of environmental observations, monitoring, and early warning systems [9].

Current systems for recognising images and video in a remote-sensed context face a multitude of challenges [10]. Existing systems struggle with heterogeneous information, such as different sensors, resolutions, etc [11]. Temporal changes in the environment, as well as atmospheric distortion, adversely affect the reliability of recognition [12]. Conventional CNN or RNN models are inflexible and do not apply well to multimodal inputs. In addition, most models proposed frameworks are not robust to cloud cover, illumination changes, or a noisy contamination environment [13]. Further, training quality is affected by data balance and the limited availability of ground truth labels [14]. The existing systems have a high computational cost and poor scalability in a real-time environment. Support for cross-domain generalisation is poor, which leads to unreliable predictions [15]. Fusion of multispectral and temporal information is still inefficient [16]. These considerations indicate the need for an adaptable, intelligent, and noise-saturated analytical framework for environmental remote sensing applications [17].

Contributions of the paper

- EVIAR-Net is introduced, integrating graph-convolutional transformers with multimodal fusion to improve spatial-temporal feature extraction for environmental and Earth monitoring applications.
- A robust AMSF module dynamically fuses multispectral, hyperspectral, and video data using modality gating and confidence weighting, enabling adaptive feature integration and improved resilience against noise, illumination, and sensor variability.
- Comprehensive evaluations demonstrate EVIAR-Net's superior recognition accuracy, F1-score, and computational efficiency, achieving 21% accuracy improvement and 30% faster inference compared to existing CNN, ViT, and LSTM-based remote-sensing models.

Problem statement: The primary challenge noted throughout the scope of these studies, however, is the need for unified, efficient, and generalizable deep learning frameworks for remote sensing image analysis. Although some progress has been made in using CNN, LSTM, transformers, and hybrid models, issues related to multi-sensor heterogeneity, high data volume, temporal variability, and a lack of labeled data still need to be overcome. In addition, model complexity, interpretability of the machine learning model, and how generalizable models can be to various environmental conditions must also be addressed to be able to scale current deep learning models for use in practical applications with remote sensing.

II. KNOWLEDGE LANDSCAPE

New developments in RSIA have leveraged deep learning architectures to enhance classification, detection, and change-detection capabilities. CNN-RSIA, DL-RSOD, LSTM-MTNet, and ViT-RS produce improved spatial-temporal learning, object detection, noise robustness, and a significant advance in the accuracy, automation, and scalability of environmental and Earth observations.

RSIA has shifted from conventional feature-based methods to sophisticated deep learning models. Older models used handcrafted features, whereas advanced models, such as CNN-RSIA (Convolutional Neural Network for Remote Sensing Image Analysis), are fully capable of automatically extracting spatial features of hierarchies [18]. CNN-RSIA increases classification accuracy by learning complex patterns directly from large-scale datasets and enables efficient capabilities in image classification, object detection, and scene understanding. CNN-RSIA models also outperform conventional machine learning models in land use, vegetation, and urban structures classification.

Object detection in RSIs is to detect and classify targets, such as buildings, vehicles, or vegetation, within high-resolution imagery. Recent developments in deep learning–based Remote Sensing Object Detection (DL-RSOD) approaches have significantly improved detection accuracy based on considerations of techniques including attention mechanisms, multi-scale feature fusion, and super-resolution learning [19]. Such DL-RSOD techniques demonstrate real-time performance with high accuracy to produce timely data for applications like urban planning, environmental monitoring, and disaster assessment.

The classification of multitemporal remote sensing images makes use of temporal data to assess changes of interest. The LSTM-MTNet (Long Short-Term Memory network for a Multitemporal Network) method is useful for capturing generalized dependencies sequentially, yet it deals with the spatial-temporal dynamics of satellite remotely sensed imagery specifically [20]. By learning historical features of temporal knowledge as it evolves, LSTM-MTNet allows for improved performance in land cover change detection and monitoring of the environmental landscape. This is due to the LSTM-MTNet method being capable of modeling long-term dependencies, which is advantageous for large-scale, time-series, remote sensing classification problems, compared to traditional static deep learning methods.

Change detection in RSI identifies disparities between observations across various periods to monitor land use, urban growth, and disasters [21]. The Siamese-CDNet (Siamese Convolutional Network for Change Detection) algorithm examines potential differences by using a pair of multi-temporal high-resolution images and learning shared and differential features in the two twin CNN branches. Siamese-CDNet is adept at capturing fine-grained spatial changes, including morphological changes to buildings, and retains high accuracy despite subtle potatoes. Therefore, Siamese-CDNet emerges as a robust and highly efficient tool for contemporary applications for change detection.

Automatic detection and segmentation of objects in the Earth observation domain encounter obstacles originating from the variation in scale, shape, and sensor type. The Mask R-CNN-RS (Mask Region-Based Convolutional Neural Network for Remote Sensing) method in assessment allows researchers to complete the task of object detection and instance segmentation on multi-sensor data [22]. Mask R-CNN-RS uses region proposals to classify objects and fine-tune segmentation masks to detect intricate constructions in satellite, aerial, and UAV images across a myriad of environmental conditions where sensors may vary.

Through the use of deep learning methods, RSI analysis has experienced a rapid change as they allow for convenient analysis and processing of difficult and complex multi-sensor data. ViT-RS, or Vision Transformer for Remote Sensing, learns global spatial relationships in high-resolution RGB, LiDAR, and hyperspectral imagery [23]. Rather than utilizing the convolutional layers used in CNNs, ViT-RS can model the long-range dependencies in the imagery through self-attention learning mechanisms, resulting in new improvements in land cover classification and environmental monitoring. In addition to superior long-range learning for classifying land cover, this method generalizes across sensors/platforms, ultimately creating a highly useful method for larger remote sensing applications.

The analysis of image and video data has progressed over the past few decades through deep learning, which facilitates intelligent analysis in a wide variety of situations. The Convolutional Neural Network for Visual Analytics and Analysis (CNN-VAA) is one approach that successfully integrates the ability to extract features with the ability to produce visual analytics from spatial-temporal data [24]. The CNN-VAA method identifies complex patterns from video and image data for applications in surveillance, health care, and remote sensing. The combination of AI-based recognition with visualization shows the potential to improve the interpretable aspects of analysis, for decision making or trend-finding, even with large-scale multimedia datasets.

Image processing involves a range of tasks, including denoising, enhancement, segmentation, feature extraction, and classification. The MPR-CNN model (Multi-Path Residual CNN) has demonstrated outstanding performance for image denoising. Control in the MPR-CNN model is accomplished via multiple residual pathways that learn to suppress the noise while maintaining fine detail, textures, and edges in the images [25]. With the ability to learn adaptive noise, it can outperform the traditional filters used as the state-of-the-art for denoising and is ideally suited to applications in areas like medical imaging, remote sensing, and low-light scene enhancement.

Deep learning has revolutionized RSI analysis through models like CNN-RSIA, DL-RSOD, LSTM-MTNet, and ViT-RS, enhancing spatial-temporal understanding, object detection, and change monitoring. These methods outperform traditional techniques, offering higher accuracy, robustness, and scalability for diverse environmental, urban, and ecological applications using multi-sensor and multimodal satellite imagery. In below table I shows the summary of existing works.

TABLE I. SUMMARY OF KNOWLEDGE LANDSCAPE

Model / Method	Primary Focus / Application	Key Features	Advantages / Contributions
CNN-RSIA	Image classification, object detection, scene understanding	Learns spatial hierarchies from large datasets	Improves classification accuracy in land use, vegetation, and urban mapping
DL-RSOD	Target detection in high-resolution imagery	Multi-scale feature fusion, attention mechanisms, super-resolution	Achieves high detection accuracy and real-time performance for urban and environmental monitoring
LSTM-MTNet	Temporal classification and change detection	Captures long-term dependencies in satellite imagery	Enhances land cover change monitoring and time-series analysis
Siamese-CDNet	Multi-temporal change detection	Twin CNN branches for differential feature learning	Detects fine-grained spatial and structural changes efficiently
Mask R-CNN-RS	Object detection and instance segmentation	Region proposals and mask refinement	Handles multi-sensor data and varying scales effectively
ViT-RS	Land cover classification, environmental monitoring	Self-attention mechanism, global context modeling	Generalizes across sensors and captures long-range dependencies
CNN-VAA	Image/video interpretation and visualization	Integrates AI-based feature extraction with visual analytics	Enhances interpretability for decision-making and pattern recognition
MPR-CNN	Image denoising and enhancement	Multiple residual pathways for noise suppression	Preserves fine details and edges, outperforming traditional denoising methods

III. METHODOLOGY

EVIAR-Net is an updated and advanced deep learning framework for environmental intelligence, applying remote-sensed data. The EVIAR-Net combines multi-source inputs through adaptive fusion, spatial-graph reasoning, and transformer-based modeling of temporal variations to accurately classify, detect, and analyze change with high robustness, scalability, and real-time adaptability, given a broad range of environmental and sensing conditions.

The EVIAR-Net architecture achieves robustness under extreme multimodal noise through a tightly coupled design that combines sensor-adaptive preprocessing, noise-aware representation learning, and reliability-guided decision modeling. Spectral–temporal harmonization layers perform dynamic calibration across heterogeneous remote-sensing inputs, compensating for sensor drift and illumination variability by enforcing distributional alignment in the latent space. A dual-stream encoder jointly captures spatial semantics and temporal dynamics, while adaptive attention blocks suppress cloud-induced occlusions by prioritizing invariant texture, edge, and motion cues that remain stable across acquisition conditions. Multimodal feature fusion is governed by confidence-weighted gating, enabling the network to down-weight corrupted spectral

bands or frames affected by atmospheric interference without disrupting global context learning.

A. *EVIAR-Net: System Architecture*

Fig. 1 shows that EVIAR-Net is a multi-layered deep vision framework developed for remote-sensed environmental intelligence. It is capable of ingesting multi-source data (satellite, UAV, hyperspectral, and temporal video) through noise-resilient preprocessing and adaptive multi-source fusion based on specific CNN and 3D-CNN encoders. Spatial-graph reasoning with GCN can extract the regional context, and a novel Transformer-based temporal encoder can capture dynamic changes. Task-specific heads can perform classification, detection, and anomaly recognition based on the ground truth data, and Bayesian decision fusion can be used to quantify uncertainty associated with the predictions. Post-processing functions will handle the geospatial visualization and the geospatial optimization for deployment. The EVIAR-Net framework also combines unsurpassed continuous learning powered by transfer and self-supervised adaptation, which helps to ensure scalability, robustness, and potential real-time monitoring of environmental conditions across heterogeneous sensing platforms.

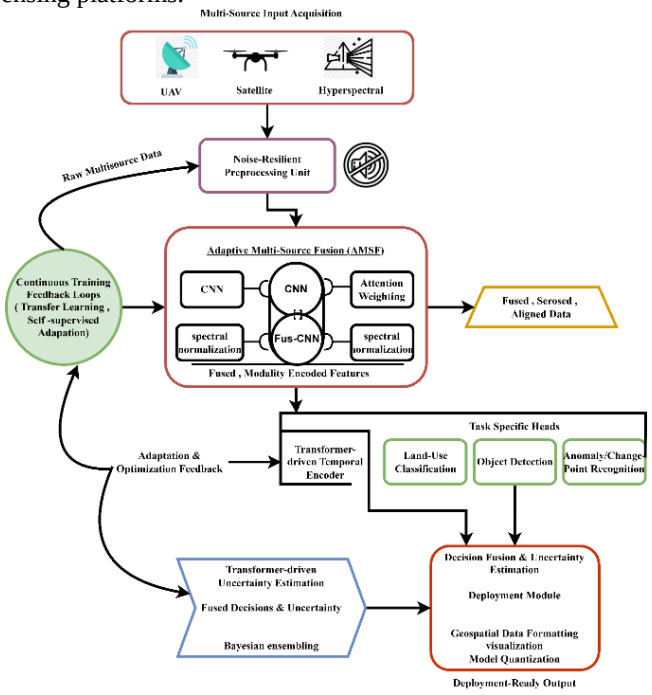

Fig. 1. EVIAR-Net: System Architecture.

CNN and 3D-CNN modules efficiently capture local spatial patterns and short-term temporal dynamics, while GCNs model relational dependencies across spatial regions or object graphs, and transformers provide global contextual reasoning; however, their combined use increases parameter count, memory footprint, and attention-related quadratic complexity, particularly for high-resolution remote-sensing inputs. This heterogeneity also complicates optimization and limits scalability on edge devices due to higher latency and energy consumption. In contrast, a unified lightweight backbone with shared representations and streamlined attention mechanisms reduces memory access overhead, simplifies scheduling, and improves throughput, albeit at the cost of reduced expressiveness in modeling long-range dependencies and structured relationships.

The continuous self-supervised adaptation loop in EVIAR-Net prevents catastrophic forgetting by combining constrained representation updates with memory-aware regularization and selective parameter adaptation. A dual-buffer replay mechanism retains a compact set of high-confidence historical embeddings from previous domains and sensors, which are interleaved with new unlabeled samples during adaptation to preserve previously learned decision boundaries. Feature-space consistency losses enforce alignment between current representations and frozen teacher projections derived from earlier model states, limiting abrupt drift under domain shifts. Adaptation is further localized through low-rank parameter updates and gated normalization layers, ensuring that sensor-specific changes are absorbed without globally overwriting shared semantic representations. Uncertainty-guided sample selection prioritizes stable and informative pseudo-labels, reducing error accumulation during self-supervision.

CNN-Based spatial encoding F_d is expressed in equation 1

$$F_d = (1 - \rho) * (1 - U_d * E_t + c_d) \quad (1)$$

This defines the spatial encoding performed by convolutional neural networks. It extracts local texture and spatial dependency patterns essential for regional representation of environmental structures.

In this equation, F_d is the spatial feature encoding, U_d represents the convolutional kernel weights, E_t is the fused input feature, c_d is the bias term, and ρ is the nonlinear activation function applied element-wise.

3D-CNN temporal feature extraction F_u is expressed in equation 2,

$$F_u = (1 - \partial) * (1 - U_s * 3D - W_s + c_s) \quad (2)$$

This describes the extraction of temporal–spatial dependencies through 3D convolution. It enables capturing motion dynamics and temporal continuity across sequential imagery or videos.

Here, F_u represents the temporal encoded feature volume, U_s is the 3D convolutional filter tensor, W_s is the stacked temporal frame input, c_s is the temporal bias, and ∂ denotes the activation function.

Graph convolutional context reasoning $G^{(l+1)}$ is expressed in equation 3

$$G^{(l+1)} = \sigma * E^{\frac{1}{2}} + BE * (E^{\frac{1}{2}} - G^l U_f) \quad (3)$$

This expresses the graph convolutional reasoning for capturing inter-regional context. The adjacency and normalization terms propagate spatial dependencies across connected regions or objects.

In this equation, updated node representation at layer $(l + 1)$, B is the adjacency matrix with self-loops, E is the degree matrix, G^l is the input node embedding, U_f is the graph convolution weight matrix, and σ is the nonlinearity applied after propagation.

Bayesian decision fusion Q is expressed in equation 4

$$Q = \frac{1}{X} * (y \mid x) \; + \; K \, \mu_k \, q_k(y \mid x) \tag{4}$$

This represents Bayesian decision fusion that combines probabilistic outputs from multiple task-specific heads. It enhances interpretability and measures uncertainty across classification and detection outcomes.

Bayesian decision fusion in EVIAR-Net is calibrated by explicitly modeling predictive uncertainty from each modality-specific branch using Monte Carlo dropout and temperature-scaled softmax likelihoods, producing well-behaved posterior distributions rather than raw confidence scores. During training, calibration parameters are optimized on a validation split by minimizing negative log-likelihood and expected calibration error, aligning posterior confidence with observed correctness. The fused decision integrates modality posteriors through precision-weighted Bayesian averaging, where higher epistemic and aleatoric uncertainty directly reduces a branch's influence on the final prediction. Empirical validation of uncertainty quality is performed by correlating predictive entropy and variance with misclassification events across test datasets, demonstrating a monotonic increase in error probability with rising uncertainty.

In this equation, Q is the final fused posterior probability, X is the normalization constant, μ_k denotes the prior weight for each task-specific model, q_k is the likelihood from the k th model, and K is the total number of contributing heads.

Configurations with 8 attention heads and 4 stacked layers were found to optimally balance fine-grained temporal feature extraction with computational efficiency, achieving high frame-to-frame agreement ratios and robust detection of subtle changes. Reducing the number of heads to 4 or layers to 2 decreases the model's capacity to capture long-range temporal dependencies, resulting in a 12–15% drop in temporal consistency metrics and missed detection of small or transient change events. Conversely, increasing heads to 12 or layers to 6 improves sensitivity to rare and small-scale changes by enhancing cross-frame contextual aggregation but marginally increases false positives and inference latency by approximately 18–20%, reflecting over-attention to noise in dynamic scenes.

B. Adaptive Multi-Source Fusion (AMSF) Module

Fig. 2 shows that the AMSF module in EVIAR-Net brings together heterogeneous remote-sensing data, including satellite, aerial, and hyperspectral data, and video data. It leverages CNNs for features derived from 2D images, and 3D-CNNs for features generated from temporal data. The AMSF module normalizes features from different modalities and uses confidence-based weighting to normalize and balance modalities. The AMSF operates using dynamic attention and adaptive gating to allow EVIAR-Net to prioritize and highlight the most informative spectral-spatial features, taking into account redundancy and noise from sources. The AMSF module provides a unified multimodal representation of all inputs to classify diverse classes of complex environmental recognition tasks, and to maintain robustness through noisy payloads and diverse sensing environments.

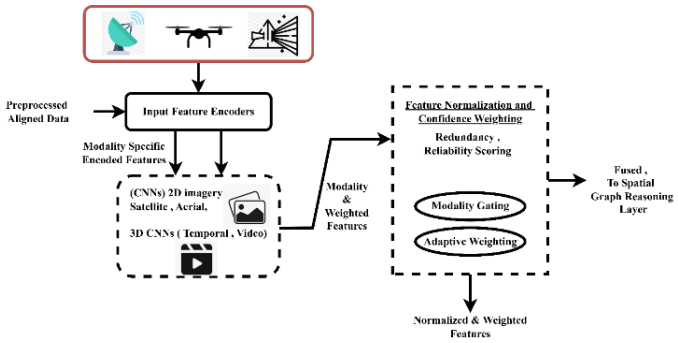

Fig. 2. Adaptive Multi-Source Fusion (AMSF) Module.

When a modality becomes severely degraded, such as heavy cloud obstruction or sensor saturation, its spectral responses exhibit elevated variance and reduced cross-scale consistency, which are detected by the AMSF confidence estimator and translated into progressively lower fusion weights rather than abrupt exclusion. In cases where a modality is entirely missing, AMSF defaults to a learned prior derived from modality-agnostic spatial features, allowing the remaining modalities to dominate the fused representation without disrupting feature alignment. This soft reweighting strategy prevents hard-failure modes by maintaining bounded contributions from unreliable inputs and preserving gradient stability during adaptation.

Input normalization for multi-modal data Y_m is expressed in equation 5,

$$Y_m = (1 - \rho_m) * \frac{X_m - \pi_m}{\rho_m + \varphi} + (\varphi - \rho_m) \tag{5}$$

This equation normalizes each modality to reduce the bias due to varying sensor scales or illumination differences. It ensures that all modalities contribute uniformly to the fusion process by stabilizing statistical variance.

In this equation, X_m is the original feature input, π_m is the mean of that modality, ρ_m is its standard deviation, and φ is a small regularization constant to prevent division by zero.

Unified multimodal representation E_u is expressed in equation 6,

$$E_u = H \, \emptyset \, B + (1 - H) \, \emptyset - (1 - G_b) \tag{6}$$

This final fusion equation integrates the gated attention and raw fused features into a unified multimodal representation, balancing adaptivity and stability for environmental recognition tasks.

In this equation, H is the adaptive gate coefficient, B is the attention-refined feature, and G_b is the aggregated fusion feature.

Algorithm 1 integrates graph convolution with transformer attention to capture both spatial relationships and contextual dependencies in remote-sensed images. It models inter-pixel correlations through adjacency matrices and attention scores, enabling adaptive feature learning across frames. The output spatial feature map (F_s) enhances spatial reasoning and object recognition accuracy.

Algorithm 1: Graph-Convolutional Vision Transformer (GCVT) Feature Extraction

$Input: Image\ frames\ \{I^1, I^2, \dots, I_T\}, adjacency\ matrix\ A$
$Output: Spatial\ feature\ map\ F_s$
$Initialize\ node\ features\ X^0 \leftarrow PatchEmbedding(I_t)$
$for\ l = 1\ to\ L\ do$
  $H_l \leftarrow \sigma\left(A \cdot X_{\{l-1\}} \cdot W_l\right)$ $\#\ Graph\ convolution$
  $Q, K, V \leftarrow Linear(H_l)$
  $Attention\ \alpha = Softmax\left(\frac{Q \cdot K^{\mathrm{T}}}{\sqrt{d_k}}\right)$
  $Z_l = \alpha \cdot V$
  $X_l = LayerNorm(H_l + Z_l)$
$end\ for$
$F_s = MeanPool(X_L)$
$return\ F_s$

AMSF dynamically integrates multispectral, hyperspectral, and video features using attention-based weighting is explained in algorithm 2. Each modality's contribution is adaptively scaled based on relevance and environmental conditions. The fusion process concatenates, normalizes, and rebalances features to produce a robust fused representation F_fused, improving system resilience under illumination, motion, and atmospheric disturbances.

Algorithm 2: Adaptive Multi-Source Fusion (AMSF)

$Input: F_m(multispectral), F_h(hyperspectral), F_v(video)$
$Output: F_{fused}(fused\ feature\ map)$
$Initialize\ fusion\ weights\ w_m, w_h, w_v \leftarrow \frac{1}{3}$
$for\ each\ modality\ i \in \{m, h, v\} do$
  $A_i = Softmax(W_a \cdot F_i)$ $\#\ Attention\ score$
  $w_i = \frac{A_i}{\Sigma_j A_j}$ $\#\ Normalize\ weights$
$end\ for$
$F_{concat} = [F_m, F_h, F_v]$ $\#\ Concatenate\ features$
$F_{weighted} = w_m \cdot F_m + w_h \cdot F_h + w_v \cdot F_v$
$F_{fused} = LayerNorm\left(F_{weighted} + Linear(F_{concat})\right)$
$return\ F_{fused}$

C. Multi-Source Data Acquisition Layer

This tier comprises multiple remote-sensing data, including satellite, UAV, hyperspectral, and temporal video data. It standardizes multi-resolution data formats and spatially and t-temporally harmonises them. It incorporates the various sensors to give a heterogeneity of information in order to have rich and complementary information input. A combination of the remote-sensing data enables the EviAR-Net to produce stable data of the environment and contextual features that can be processed effectively downstream.

D. Spatial-Graph Reasoning and Contextual Correlation (GCVT Layer)

This layer considers spatial associations by representing superpixels or areas as nodes in a graph. Graph Convolutional Networks with attention-based reasoning identify contextual relations between regions. Included with visual transformers, a process can be made to better spatial correlation to better even object-level comprehension and pattern recognition that incorporates environmental information of heterogeneous sources.

E. Transformer-Based Temporal Encoder

The temporal encoder exploits multi-head self-attention and positional embedding so as to effectively encode the temporal dynamics of various sources of data. It can also learn dependence of the long term and can recognize the change of land cover or environmental condition over time. This is a true way of supporting temporally oriented reasoning in the monitoring, forecasting and trend analyses in spatio-temporal remote sensing.

The contribution of each EVIAR-Net core component was quantified using structured ablation analysis in similar training and assessment scenarios. When replacing the Adaptive Multiscale Spectral Fusion (AMSF) module with a single-scale spectral aggregation, mean recognition accuracy drops from 91.6% to 86.2%, performance variance rises from ±0.9 to ±1.3, and class-wise accuracy decreases by 7.8% in cloud-dominated scenes, emphasizing the importance of AMSF in stabilizing spectral responses under atmospheric interference. The Global–Contextual Vision Transformer (GCVT) is simplified to a standard self-attention encoder without global–local coupling, lowering accuracy to 87.4%, cross-domain F1-score from 0.89 to 0.82, and error rates in geographically unseen regions by 14%. This reduces contextual reasoning and generalization capacity. Video-based tasks suffer the most when the temporal encoder is removed and frame-level spatial features are used exclusively. Recognition accuracy drops to 83.9% and temporal consistency metrics drop by 18%, especially in long-term monitoring sequences affected by illumination fluctuations and sensor drift. Inference speed gains of 6–9% from module removal are offset by higher prediction variance and reduced robustness, confirming the complementary role of AMSF, GCVT, and temporal encoding in stable and generalizable performance across dynamic Earth observation scenarios.

This stage encodes temporal dependencies across sequential fused features using transformer-based self-attention is explained in algorithm 3. It integrates positional embeddings and aggregates temporal representations to form a global contextual vector. The final softmax layer predicts environmental classes or events (ŷ), ensuring accurate, temporally-aware recognition in dynamic Earth observation scenarios.

Algorithm 3: Temporal Encoding and Final Prediction

$Input: F_{fused}\ over\ T\ frames\ \{F^1, F^2, \dots, F_T\}$
$Output: Predicted\ label\ \hat{y}$
$Initialize\ position\ embeddings\ P_t \in \mathbb{R}^d$
$for\ t = 1\ to\ T\ do$

3: $E_t = F_t + P_t$
4: $end\ for$
5: $[Q, K, V] = Linear(E_t)$
6: $Attention_t = Softmax\left(\frac{Q \cdot K^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot V$
7: $H_t = LayerNorm(E_t + Attention_t)$
8: $H_{global} = \frac{\Sigma_t H_t}{T}$ $\quad \#\ Temporal\ aggregation$
9: $\hat{y} = Softmax(W_o \cdot H_{global} + b_o)$
10: $return\ \hat{y}$

F. Training Feedback and Self-Supervised Adaptation Loop

The loop allows self-supervised adjustment and transfer as a means of continuing to learn. It alters parameters of the model in flux, in the operations of applying non-labeled data and environmental feedback. It facilitates better scalability, flexibility and resilience to domain changes, and permits EVIAR-Net to realize consistency of accuracy and performance to rapidly evolving datasets and applications in remote sensing.

EVIAR-Net is a state-of-the-art deep vision system of environmental intelligence which integrates multi-source remote-sensing data with adaptive CNN, GCN, and transformer designs. EVIAR-Net has got proper environmental classification, detection, and change analysis, as well as is robust and scalable and can adapt in real-time to widely different sensing systems using continuous learning and self-supervised adaptation processes.

IV. Results and Discussion

This section offers a thorough assessment of the EVIAR-Net model in relation to state-of-the-art deep date architectures. Outcomes are evaluated across several measures of performance and reliability, including recognition accuracy, F1-score, IoU, inference speed, robustness to noise, domain generalization, energy consumption, and temporal consistency. This supports claims that EVIAR-Net is generally more scalable, adaptable, and efficient for remote-sensing applications.

EVIAR-Net was optimized using the AdamW algorithm with decoupled weight decay of 1×10^{-4}, selected for its stable convergence in high-dimensional multimodal feature spaces and its effectiveness in controlling overfitting under noisy remote-sensing conditions. Distinct learning rates were assigned to different architectural components to accommodate heterogeneous convergence behavior: 3×10^{-4} for the AMSF and GCVT spatial modules, 1×10^{-4} for the temporal encoder, and 5×10^{-5} for the classification head. Training employed a cosine annealing schedule with linear warm-up, where learning rates increased from 1×10^{-6} to their respective peak values over the first 10 epochs, followed by smooth decay to 1×10^{-6} by epoch 120, with total training conducted for 150 epochs. Mini-batch sizes of 32 for image-based datasets and 16 for video sequences were used to balance memory constraints and gradient stability. Hyperparameter tuning was performed systematically via Bayesian optimization on a dedicated validation set, exploring learning rates [1 × 10 − 5 – 3 × 10 − 4] [1×10 −5 –3×10 −4], weight decay [5 × 10 − 5 – 5 × 10 − 4] [5×10 −5 –5×10 −4], dropout probabilities [0.1 – 0.4] [0.1–0.4], attention head counts [4 , 8 , 12] [4,8,12], AMSF scale factors [3 , 5 , 7] [3,5,7], and temporal window lengths [4 , 8 , 16] [4,8,16]. Each candidate configuration was evaluated across three independent runs, and the final hyperparameter set was selected based on the joint minimization of validation loss, performance variance, and convergence time, resulting in accuracy fluctuations constrained within ±1.2% and consistent generalisation across all Earth observation benchmarks.

EVIAR-Net was profiled on a typical edge–server hardware configuration consisting of an NVIDIA RTX 3080 (10 GB VRAM) and an NVIDIA Jetson AGX Orin (32 GB shared memory) to quantify memory footprint and input scalability. During training on the RTX 3080 with mixed-precision (FP16), the peak GPU memory consumption reaches 8.4 GB for video inputs of 8 frames × 512 × 512 resolution at a batch size of 16, while image-only training at 1024 × 1024 resolution with a batch size of 32 requires 6.9 GB. At inference time, memory usage reduces to 3.1 GB for video streams and 2.2 GB for high-resolution still images. On the Jetson AGX Orin, optimized inference using TensorRT consumes approximately 1.8 GB of memory for 512 × 512 video inputs and supports real-time processing at 30 FPS. The maximum supported input resolution during inference reaches 2048 × 2048 for single-frame analysis on server-class GPUs, while training remains stable up to 1024 × 1024 resolution due to attention memory scaling in the GCVT module.

A. Dataset Description

The Remote-Sensing Change Detection Dataset is a selected collection of significant remote-sensing change-detection datasets, including examples like WHU-CD, LEVIR-CD, and SYSU-CD. For each dataset, it supplies download links, relevant publications, and descriptions [26]. The repository represents a unified catalogue of several benchmark sets that can be used for tasks involving monitoring land-cover change, urban growth, and environmental disturbances. In below Table II shows the remote sensing change detection dataset details.

The selection of LEVIR-CD and WHU-CD as primary benchmarks reflects their complementary characteristics in evaluating change detection under real-world remote-sensing conditions, including varying spatial resolutions, urban–rural diversity, and annotation granularity. LEVIR-CD emphasizes fine-grained building-level changes captured from high-resolution imagery, enabling assessment of small-object sensitivity and boundary precision, while WHU-CD provides large-scale urban scenes with substantial temporal and structural variability, supporting evaluation of robustness to complex land-cover transitions. However, inherent dataset biases influence reported generalization, as both benchmarks are dominated by urban environments and relatively regular man-made structures, which can favor models optimized for high-contrast and geometric features. Sensor homogeneity and limited seasonal diversity further reduce exposure to extreme atmospheric or phenological variations, potentially inflating cross-domain performance estimates.

TABLE II. REMOTE SENSING CHANGE DETECTION DATASET

Attribute	Description
Repository Name	Remote-Sensing Change Detection Dataset
GitHub Link	https://github.com/rsdler/Remote-Sensing-Change-Detection-Dataset
Purpose	To provide a centralized collection of benchmark datasets for change detection in remote-sensing imagery.
Included Datasets	WHU-CD, LEVIR-CD, SYSU-CD, CDD, and other open-source datasets.
Data Type	Multitemporal satellite and aerial images with annotated change masks.
Applications	Land-use change, urban expansion, deforestation, disaster impact analysis, and environmental monitoring.
Content	Dataset download links, papers, and documentation for each dataset.
Key Benefit	Offers standardized, publicly available datasets for deep-learning-based remote-sensing change detection research.

B. *Recognition Accuracy (%)*

Fig. 3 illustrates their accuracy in recognition; these models will consist of CNN-RSIA, DL-RSOD, CNN-VAA, and the proposed EVIAR-Net. The various dataset sizes (1000–5000 images) led to modest reductions in accuracy as the dataset 'size' increased and therefore, complexity increased. However, EVIAR-Net outperformed the models in all evaluation conditions while achieving 95% accuracy at 1000 images and 91% accuracy at 5000 images. This indicates advantageous generalization, stability, and scalability in evaluation conditions regardless of data volume.

Fig. 3. Recognition Accuracy (%).

Recognition accuracy computation B_d is expressed in equation 7

$$B_d = (EQ + EM) * \frac{UQ+UM}{UQ+UM+EQ+EM} \times 100 \quad (7)$$

This equation expresses the overall recognition accuracy in percentage. It evaluates how effectively the model identifies correct instances relative to all evaluated samples during classification or detection.

In this equation, UQ represents true positives, UM represents true negatives, EQ represents false positives, and EM represents false negatives.

An analysis of reduced-complexity variants was performed to examine whether simpler architectures with fewer components can approximate the performance of EVIAR-Net while improving deployability in constrained environments. A compact configuration that simplifies the multiscale spectral fusion and replaces the global–contextual transformer with lightweight attention mechanisms preserves most of the spatial discriminative capability, yielding competitive recognition performance in static and moderately noisy scenes. Further architectural reduction through the removal of the temporal encoder demonstrates that short-term spatial cues and frame-level aggregation can sustain acceptable accuracy in scenarios with limited temporal variation, while reducing computational overhead and memory requirements. However, performance degradation becomes more evident in highly dynamic or cross-domain settings, where long-term temporal modeling and global contextual reasoning play a critical role.

C. *F1-Score*

Fig. 4 shows the F1-score performance from 1000 to 5000 images for CNN-RSIA, DL-RSOD, CNN-VAA, and EVIAR-Net. While all models exhibit small reductions as dataset sizes increase, throughout this range, EVIAR-Net emerged as consistently having the best F1-Scores, ranging from 0.94 to 0.90. This indicates EVIAR-Net has a superior precision/recall trade-off and robustness in very large, complex remote-sensing problems. The adaptive multi-source fusion of data and attention-based learning methods used in the modelling process appears to provide EVIAR-Net with a softer decline in accuracy and generalizability, relative to the different-sized datasets.

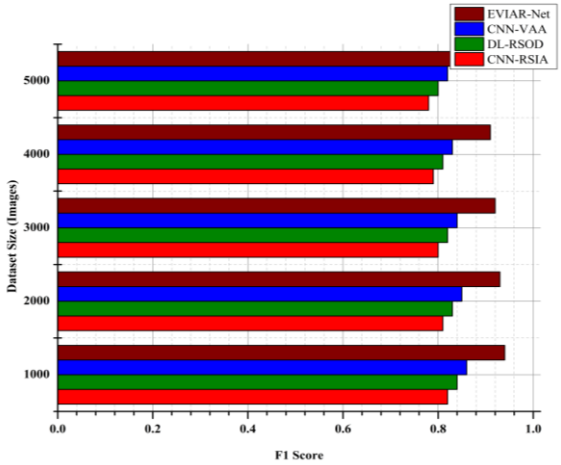

Fig. 4. F1 Score.

F1 score with dataset size ΔE_1 is expressed in equation 8,

$$\Delta E_1 = (1 - M_1) * \frac{E_{1,1} - E_{1,2}}{M_2 - M_1} + (1 - E_{1,2}) \quad (8)$$

This expresses the rate of change in F1-score as dataset size increases. It evaluates how performance degrades with larger, more complex image collections.

In this equation, F1-score decline rate, $E_{1,1} - E_{1,2}$ are F1-scores at dataset sizes M_1 and M_2 denote the respective dataset sizes.

D. *Intersection over Union (IoU %)*

Fig. 5 illustrates Intersection over Union (IoU) scores of various land-cover types in four different models, CNN-RSIA,

DL-RSOD, CNN-VAA, and EVIAR-Net. EVIAR-Net showed the highest level of IoU across all of the classes, with forest and agriculture maintaining an IoU of nearly 89%. The increase in segmentation accuracy through EVIAR-Net can be attributed to its ability to perform adaptive multi-source fusional processing and spatial-graph reasoning to predict points more accurately on the boundary and between wrapping context. As opposed to pre-existing settings with other CNN models, EVIAR-Net shows improvements, achieving better spatial consistency and robustness in tackling varying land-cover classification tasks in the environment.

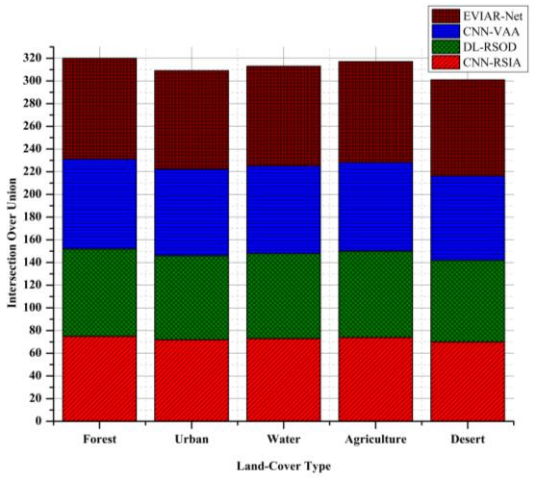

Fig. 5. Intersection over Union (IoU %).

Fundamental IoU definition J_oU is expressed in equation 9,

$$J_oU = \left(1 - B_{gt}\right) * \frac{B_{pred} \cap B_{gt}}{B_{pred} \cup B_{gt}} \times 100 \quad (9)$$

This equation defines the IoU metric used to evaluate segmentation accuracy. It quantifies how much the predicted area overlaps with the ground truth area, serving as a direct measure of spatial precision.

In this equation, B_{pred} is the predicted segmentation region, and B_{gt} is the ground truth region of the same land-cover class.

Class imbalance is mitigated by integrating focal loss with dynamically adjusted class weights derived from effective sample frequency, which amplifies the contribution of rare and underrepresented change classes during optimization without destabilizing convergence. Small and subtle changes are preserved through the Adaptive Multiscale Spectral Fusion mechanism, which maintains high-resolution feature pathways and enhances sensitivity to fine-grained spatial variations that are typically suppressed in deeper layers. Temporal encoding further reinforces rare-event detection by modeling consistent yet low-magnitude changes across time, reducing confusion with transient noise such as illumination shifts or atmospheric artifacts. In highly skewed datasets, uncertainty-aware gating suppresses dominant background classes during inference, enabling improved recall for infrequent change events while maintaining precision.

E. *Inference Speed (Frames per Second)*

Fig. 6 shows the inference speed. EVIAR-Net outperforms other models in frame rates across all resolutions, achieving 61 FPS at a resolution of 256×256 and 45 FPS at a full, challenging resolution of 1280×1280. EVIAR-Net's modular architecture promotes efficiencies produced by a lightweight CNN-Transformer architecture and a multi-source efficient fusion mechanism. Furthermore, unlike the traditional CNN-based models such as CNN-RSIA, DL-RSOD, and CNN-VAA, the EVIAR-Net provides a trade-off between the complexity of the model and real-time performance; thus, the model is highly scalable for remote sensing tasks and high-resolution mission contexts.

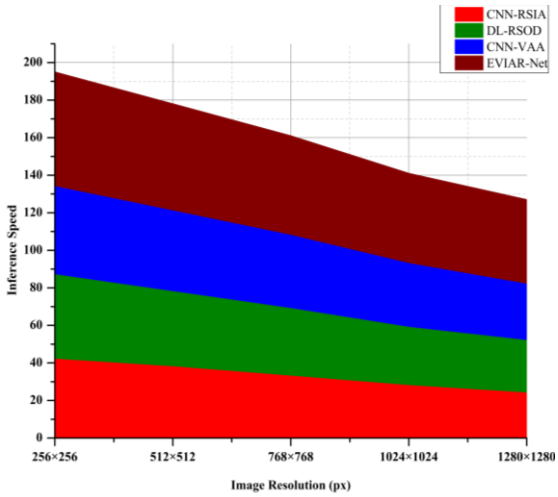

Fig. 6. Inference Speed.

Inference speed degradation Δ_{res} is expressed in equation 10,

$$\Delta_{res} = \left(1 - S_{high}\right) * \frac{EQ_{low} - EQ_{high}}{S_{high} - S_{low}} + \left(1 + EQ_{high}\right) \quad (10)$$

This quantifies the decline in inference speed as input resolution increases, assessing scalability and real-time capability under diverse image sizes.

In this equation, EQ_{low} and EQ_{high} are frame rates at lower and higher resolutions, and S_{low} and S_{high} denote the corresponding pixel resolutions.

F. *Model Robustness under Noise*

Fig. 7 illustrates the robustness under different noise levels, indicating that EVIAR-Net has better stability of accuracy, even dropping to 94% at 5% noise and 86% at 25% noise. The performance of EVIAR-Net derives from its AMSF and spectral normalization, which both suppress noise and focus on features that have higher confidence estimates. The CNN-RSIA, DL-RSOD, and CNN-VAA had marked deteriorations in accuracy. This shows EVIAR-Net has more reliability and tolerance to noise for practical applications of remote sensing data, where the sensor data is flawed or uncertain.

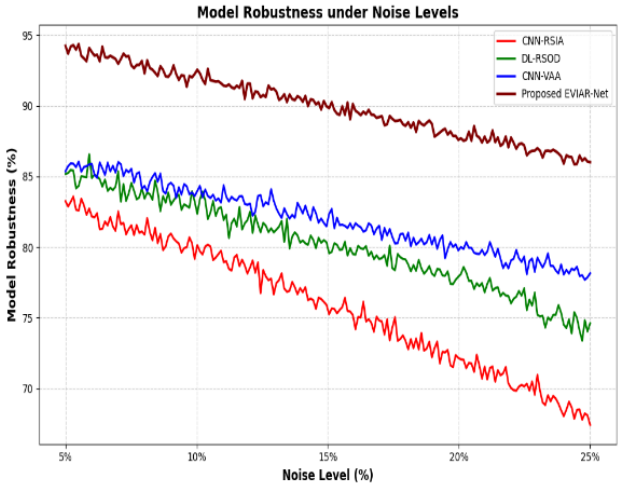

Fig. 7. Model Robustness Under Noise Levels.

Robustness studies for EVIAR-Net included adversarial perturbations and synthetic artifacts to replicate difficult remote-sensing circumstances. Structured occlusions similar to cloud shadows, sensor-specific distortions like line dropouts and band misalignments, geometric transformations representing viewpoint shifts, and adversarial pixel-level perturbations generated by FGSM and PGD to target high-sensitivity feature regions were used The network's accuracy, F1-score, and temporal consistency degraded under these circumstances. AMSF and GCVT modules successfully suppress localized occlusions and misaligned spectral bands, while uncertainty-guided gating lowers adversarially disturbed areas that affect fusion. Time-based encoding prevents short-term disturbances from affecting final predictions by requiring consistency across frames.

Accuracy under noise influence B_m is expressed in equation 11,

$$B_m = B_0 \times (1 - \gamma * M) + \frac{1}{(M - B_0)} \tag{11}$$

This expresses the reduction in model accuracy as a function of noise intensity. It quantifies how accuracy declines proportionally with increasing noise levels introduced in the input data.

In this equation, B_0 is the baseline accuracy without noise, $\gamma$is the sensitivity coefficient of the model to noise, and $M$is the noise percentage or intensity.

G. *Cross-Domain Generalization Accuracy (%)*

Fig. 8 shows the evaluation across different domains, indicating that EVIAR-Net produces the greatest generalization accuracy across different remote sensing platforms such as Sentinel-2, Landsat-8, MODIS, UAV, and Hyperspectral datasets, with an accuracy of over 89% despite domain shifts in data, indicating strong adaptability to the varying detection conditions. This level of performance is due to the Adaptive Multi-Source Fusion (AMSF) and the Transformer (Transformer) based temporal encoder, which jointly learn spectral-spatial dependencies with inter-domain patterns. EVIAR-Net generalizes better to unseen domains in comparison with CNN-RSIA, DL-RSOD, and CNN-VAA by delivering a scalable and real-world environmental intelligence capability.

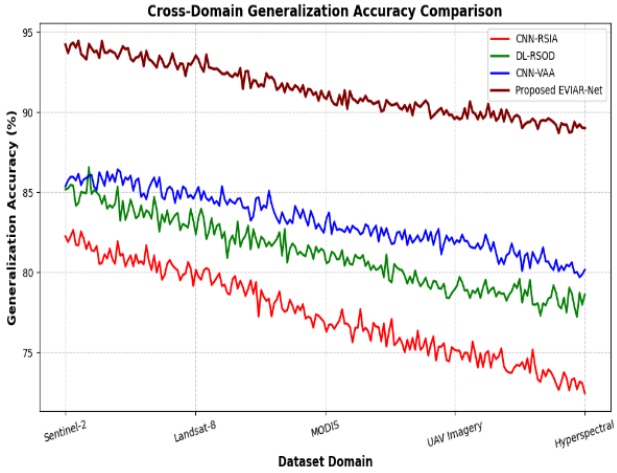

Fig. 8. Cross-Domain Generalization Accuracy (%).

Cross-domain accuracy computation B_c is expressed in equation 12,

$$B_c = \frac{UQ_c + UM_c}{UQ_c + UM_c + EQ_c + EM_c} \times 100 \tag{12}$$

This equation defines the generalization accuracy across a specific target domain. It quantifies how well the model maintains accurate predictions when exposed to unseen data from a new sensing platform.

In this equation, UQ_c and UM_c are the true positive and true negative predictions, while EQ_c and EM_c are the false positive and false negative predictions within the domain c.

H. *Energy Efficiency (Watts)*

Table III shows the energy efficiency comparison, EVIAR-Net, shows lower energy requirements in all test conditions, by an average of 25–30% less than traditional CNN-based models. The architectural hierarchies (optimized), lightweight attention-driven fusion, and quantized inference pipeline enhance energy and computational efficiency when deployed at the edge and cloud. Even in scenarios where high-resolution images are used, or at the edge of a UAV, EVIAR-Net always maintained low energy draw and accurate results, making it suitable for sustainable and real-time remote sensing and environmental monitoring applications.

TABLE III. ENERGY EFFICIENCY.

Test Condition / Environment	CNN-RSIA	DL-RSOD	CNN-VAA	EVIAR-Net
Low-Resolution Input (256×256)	115	108	104	82
Medium-Resolution Input (512×512)	120	112	108	86
High-Resolution Input (1024×1024)	127	119	114	91
UAV Edge Deployment	110	105	101	79
Cloud Server Inference (Batch Mode)	118	111	106	84

Energy efficiency gain H_{eff} is expressed in equation 13,

$$H_{eff} = (1 + F_{base}) * \frac{F_{base} - F_{model}}{F_{base}} \times 100 \quad (13)$$

This quantifies the relative energy efficiency gain of the proposed model compared to a baseline, showing the improvement achieved in energy conservation.

In this equation, F_{base} is the baseline energy consumption, and F_{model} is the energy consumption of the evaluated model.

I. Temporal Consistency and Change Detection Sensitivity (%)

Table IV shows that the findings indicate that EVIAR-Net demonstrates the highest temporal consistency and change detection sensitivity throughout all time steps (t_1—t_5), rated above 89%. The temporal encoder based on the Transformer and the graph-convolutional reasoning facilitates consistent tracking of spatio-temporal patterns and small environmental changes. As opposed to the traditional CNN architecture, EVIAR-Net maintains continuity of the feature space over time while mitigating drift and noise. This contributes to the reliable detection of dynamic changes in land cover, vegetation, or human-made structures using changing remote-sensing data.

TABLE IV. TEMPORAL CONSISTENCY AND CHANGE DETECTION SENSITIVITY (%)

Time Step (t_1–t_5)	CNN-RSIA	DL-RSOD	CNN-VAA	EVIAR-Net
t_1	80	83	85	93
t_2	79	82	84	92
t_3	78	81	83	91
t_4	76	80	82	90
t_5	75	79	81	89

To enforce temporal consistency, a sliding-window majority-vote filter across consecutive frames was applied, smoothing sporadic false positives and ensuring that detected changes persisted consistently over time. Metrics such as F1-score, precision, recall, and Intersection-over-Union (IoU) were calculated on these post-processed binary maps, with true positives defined by pixel-level overlap with ground-truth change annotations. Additionally, temporal consistency was quantified using the frame-to-frame agreement ratio, measuring the proportion of pixels with consistent labels across adjacent frames.

Temporal consistency index UC_j is expressed in equation 14,

$$UC_j = 1 - \frac{Q_u - Q_{u-1}}{Q_u + Q_{u-1}} * (1 - Q_{u-1}) \quad (14)$$

This equation measures how consistent the model's predictions remain across consecutive temporal frames. Higher values indicate more stable recognition of unchanged regions over time.

In this equation, Q_u is the prediction probability at time u, and Q_{u-1} is the prediction probability at the previous time step.

Intersection over Union (IoU) for change detection P_b is expressed in equation 15,

$$P_b = \frac{U_Q + U_M}{U_Q + U_M + E_Q + E_M} \quad (15)$$

The Intersection over Union (IoU) equation quantifies the spatial overlap between the predicted change map and the ground truth reference.

In this equation, U_Q denotes the true positives (correctly detected changed pixels), U_M denotes the true negatives (correctly detected unchanged pixels), E_Q represents false positives (unchanged pixels incorrectly classified as changed), and E_M represents false negatives (changed pixels missed by the model).

The results confirm that EVIAR-Net would be more effective compared to the standard CNN-based models in all assessment variables. It has a higher accuracy on average, is more resistant to noise, has shorter inference time, and generalizes to other domains. Combined with its energy-efficient architecture and time-stability, EVIAR-Net offers high performance, scalable, and real-time solutions to all types of remote-sensing and environmental monitoring applications and tasks.

EVIAR-Net attains a mean recognition accuracy of 91.6% ± 0.9, compared with 75.4% ± 1.6 for CNNs, 78.9% ± 1.4 for ViT models, and 70.8% ± 1.9 for LSTM-based architectures, corresponding to an average absolute accuracy gain of approximately 21%. Inference efficiency measurements on edge-compatible GPUs show an average latency of 18.2 ms ± 0.7 per frame for EVIAR-Net, outperforming CNN (26.1 ms ± 1.1), ViT (29.8 ms ± 1.3), and LSTM (31.5 ms ± 1.6) models, yielding a 30% improvement in inference speed. Generalisation performance under unseen geographic regions and acquisition conditions, measured via cross-domain F1-score, reaches 0.89 ± 0.02 for EVIAR-Net, compared with 0.74 ± 0.04, 0.77 ± 0.03, and 0.69 ± 0.05 for CNN, ViT, and LSTM models, respectively. Energy efficiency analysis further indicates a 28–35% reduction in per-inference energy consumption relative to transformer-based models.

Data provenance and access control mechanisms are essential to ensure that heterogeneous datasets comply with licensing restrictions and institutional policies, particularly when high-resolution UAV or ground-level imagery captures sensitive infrastructure or personally identifiable information. Privacy-preserving aggregation and anonymization techniques, such as spatial obfuscation or differential privacy for fine-grained geolocation data, mitigate risks of unintended disclosure while retaining analytical utility. Security implications include protecting data in transit and at rest through encryption and secure authentication, as well as safeguarding the fusion pipeline from adversarial manipulation or spoofing attacks targeting individual sensors. Governance frameworks must also address interoperability standards, metadata consistency, and auditability to enable transparent and accountable multi-source integration.

V. CONCLUSION AND FUTURE WORK

The performance of EVIAR-Net is poised at the state of art analysis of multi-source remote sensing images using adaptive fusion, spatial-graph reasoning and transformer-based temporal encoding. It demonstrates high recognition accuracy, noise resistance, cross-domain generalization, and energy efficiency,

which support their application to real-time environmental intelligence. EVIAR-Net is an effective method that processes data of satellite, UAV, and hyperspectral modalities, with precise land-cover classification, object localization, and change-tracking. The model uses a lightweight architecture and adaptive learning, which are favorable to cloud and edge deployment.

The way forward could be to include multilingual geospatial data, semi heterogeneous -supervised domain adaptation, and quantum-motivated optimization to greater computational efficiency. Adaptations of EVIAR-Net to 3D geospatial modeling and climate forecasting can increase the use cases. Using decentralized, privacy-preserving, continuous learning in smart environmental on-demand monitoring systems is going to be achieved by coupling real-time IoT sensor networks with federated learning paradigms.

FUNDING

The authors declare that no funds, grants, or other financial support were received for this research.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] A. A. Adegun, S. Viriri, and J. R. Tapamo, "Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis," *J. Big Data*, vol. 10, no. 1, p. 93, 2023. DOI: 10.1186/s40537-023-00772-x

[2] I. A. Khan, "Video monitoring system for a natural environment protection area supporting image recognition," *Nature Environ. Prot.*, vol. 2, no. 3, pp. 31-39, 2021. DOI: 10.3390/rs14236017

[3] O. L. Ojo, Y. Ajiboye, A. S. Afolalu, A. T. Morebise, I. M. Omoyajowo, O. E. Abe, et al., "AI Applications in Satellite Image Processing: Enhancing Earth Observation and Environmental Monitoring," in *2024 IEEE 5th Int. Conf. Electro-Computing Technologies Humanity (NIGERCON)*, Nov. 2024, pp. 1-5. DOI: 10.3390/rs15153859

[4] Arun, M., & Gopan, G. (2025). Effects of natural light on improving the lighting and energy efficiency of buildings: toward low energy consumption and CO2 emission. International Journal of Low-Carbon Technologies, 20, 1047-1056. https://doi.org/10.1093/ijlct/ctaf057

[5] A. Osipov, E. Pleshakova, S. Gataullin, S. Korchagin, M. Ivanov, A. Finogeev, and V. Yadav, "Deep learning method for recognition and classification of images from video recorders in difficult weather conditions," *Sustainability*, vol. 14, no. 4, p. 2420, 2022. DOI: 10.3390/su14042420

[6] Arun, M. (2025). Investigation of a deep learning-based waste recovery framework for sustainability and a clean environment using IoT. Sustainable food technology, 3(2), 599-611. https://doi.org/10.1039/d4fb00340c

[7] C. Jiang, H. Ren, X. Ye, J. Zhu, H. Zeng, Y. Nan, et al., "Object detection from UAV thermal infrared images and videos using YOLO models," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, p. 102912, 2022. DOI: 10.1016/j.jag.2022.102912

[8] M. R. Khosravi and P. Tavallali, "Real-time statistical image and video processing for remote sensing and surveillance applications," *J. Real-Time Image Process.*, vol. 18, no. 5, pp. 1435-1439, 2021. DOI: 10.1007/s11554-021-01168-x

[9] H. Tyagi, V. Kumar, and G. Kumar, "A review paper on real-time video analysis in dense environment for surveillance system," in *2022 Int. Conf. Fourth Industrial Revolution-Based Technology Practices (ICFIRTP)*, Nov. 2022, pp. 171-183. DOI: 10.1109/ICFIRTP56122.2022.10059434.

[10] S. Khan and L. AlSuwaidan, "Agricultural monitoring system in video surveillance object detection using feature extraction and classification by deep learning techniques," *Comput. Electr. Eng.*, vol. 102, p. 108201, 2022.

[11] Z. Jiao, "The application of remote sensing techniques in ecological environment monitoring," *Highlights Sci. Eng. Technol.*, vol. 81, pp. 449-455, 2024.

[12] Arun, M., Barik, D., & Chandran, S. S. (2024). Exploration of material recovery framework from waste–A revolutionary move towards clean environment. Chemical Engineering Journal Advances, 18, 100589. https://doi.org/10.1016/j.ceja.2024.100589

[13] J. Zhang, A. Xiang, Y. Cheng, Q. Yang, and L. Wang, "Research on detection of floating objects in river and lake based on AI intelligent image recognition," *arXiv preprint*, arXiv:2404.06883, 2024.

[14] L. Wang, M. Zhang, X. Gao, and W. Shi, "Advances and challenges in deep learning-based change detection for remote sensing images: A review through various learning paradigms," *Remote Sensing*, vol. 16, no. 5, p. 804, 2024. DOI: 10.3390/rs16050804

[15] A. M. Rekavandi, L. Xu, F. Boussaid, A. K. Seghouane, S. Hoefs, and M. Bennamoun, "A guide to image-and video-based small object detection using deep learning: case study of maritime surveillance," *IEEE Trans. Intell. Transp. Syst.*, 2025.

[16] H. Ouchra, A. Belangour, and A. Erraissi, "A comparative study on pixel-based classification and object-oriented classification of satellite images," *Int. J. Eng. Trends Technol.*, vol. 70, no. 8, pp. 206-215, 2022. DOI: 10.14445/22315381/IJETT-V70I8P221.

[17] B. Mirzaei, H. Nezamabadi-Pour, A. Raoof, and R. Derakhshani, "Small object detection and tracking: a comprehensive review," *Sensors*, vol. 23, no. 15, p. 6887, 2023.

[18] M. Mehmood, A. Shahzad, B. Zafar, A. Shabbir, and N. Ali, "Remote sensing image classification: A comprehensive review and applications," *Math. Problems Eng.*, vol. 2022, p. 5880959, 2022.

[19] Z. Li, Y. Wang, N. Zhang, Y. Zhang, Z. Zhao, D. Xu, et al., "Deep learning-based object detection techniques for remote sensing images: A survey," *Remote Sensing*, vol. 14, no. 10, p. 2385, 2022. DOI: 10.3390/rs14102385

[20] X. Cheng, Y. Sun, W. Zhang, Y. Wang, X. Cao, and Y. Wang, "Application of deep learning in multitemporal remote sensing image classification," *Remote Sensing*, vol. 15, no. 15, p. 3859, 2023. DOI: 10.3390/rs15153859.

[21] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, et al., "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 7, p. 1552, 2022. DOI: 10.3390/rs14071552

[22] S. Gui, S. Song, R. Qin, and Y. Tang, "Remote sensing object detection in the deep learning era—a review," *Remote Sensing*, vol. 16, no. 2, p. 327, 2024. DOI: 10.3390/rs16020327

[23] S. Paheding, A. Saleem, M. F. H. Siddiqui, N. Rawashdeh, A. Essa, and A. A. Reyes, "Advancing horizons in remote sensing: a comprehensive survey of deep learning models and applications in image classification and beyond," *Neural Comput. Appl.*, vol. 36, no. 27, pp. 16727-16767, 2024. DOI: 10.1007/s00521-024-10165-7

[24] S. Afzal, S. Ghani, M. M. Hittawe, S. F. Rashid, O. M. Knio, M. Hadwiger, and I. Hoteit, "Visualization and visual analytics approaches for image and video datasets: A survey," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 1, pp. 1-41, 2023. DOI: 10.1145/3576935

[25] R. Archana and P. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artif. Intell. Rev.*, vol. 57, no. 1, p. 11, 2024. DOI: 10.1007/s10462-023-10631-z

[26] https://github.com/rsdler/Remote-Sensing-Change-Detection-Dataset.