Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADVERSARIALLY RESILIENT REMOTE SENSING IMAGE CLASSIFICATION THROUGH CROSS-SPECTRAL ATTENTION-FUSED TRANSFORMER MODELING (CSAF-VIT)

HEMASHREE P[1], N VALLIAMMAL[2]

[1]Research Scholar, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Studies for Women, Coimbatore, India

[2]Associate Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Studies for Women, Coimbatore, India

E-mail: [1]shreehema256@gmail.com, [2]valliammal_cs@avinuty.ac.in

ABSTRACT

Adversarial threats pose significant challenges to the reliability of remote sensing image classification, particularly in defense and surveillance applications. The proposed CSAF-ViT framework introduces a cross-spectral attention-fused vision transformer designed to enhance resilience against pixel-level perturbations and preserve spatial-spectral integrity. The model begins by decomposing multi-band satellite imagery, followed by band-specific feature encoding and attention-guided fusion to capture intricate cross-band relationships. Transformer-based spatial encoding then enriches global contextual awareness, ensuring consistent representation of spectral semantics. A classification token mechanism is employed for final decision mapping, while adversarial training fortifies the model against gradient-induced distortions. The integration of spectral coherence and spatial dependency within a unified architecture enables stable feature propagation and decision reliability under manipulated inputs. CSAF-ViT offers a structured defense pipeline that emphasizes spectral alignment, attention consistency, and robust learning dynamics, establishing a reliable foundation for secure and accurate remote sensing classification in critical operational scenarios.

Keywords: *Remote Sensing Classification, Adversarial Robustness, Cross-Spectral Attention, Vision Transformer, Spectral Feature Fusion, Secure Image Recognition*

1. INTRODUCTION

Adversarial attacks introduce malicious, often imperceptible perturbations into input data to deceive machine learning models. These attacks have disrupted reliability across domains such as healthcare, cybersecurity, speech recognition, and autonomous systems [1]. Their increasing sophistication calls for defensive measures capable of ensuring model integrity. The field of adversarial attack prevention focuses on equipping models with the capability to resist, detect, or recover from such perturbations through mechanisms like adversarial training, input purification, and robust optimization [2].

Image classification has become a foundational task in computer vision, enabling recognition and categorization in numerous real-world settings. Within this scope, remote sensing image classification is a specialized branch concerned with analyzing satellite and aerial imagery to identify various land cover types and environmental patterns [3]. These applications are integral to urban development, agriculture, climate monitoring, and resource management [4].

Remote sensing imagery is characterized by its spectral richness and high-resolution spatial patterns. The complexity of these images makes classification models vulnerable to even small adversarial perturbations [5]. Perturbed pixels can cause misclassification of critical classes such as water bodies, vegetation, and urban infrastructure, thereby undermining decision-making processes dependent on remote sensing data [6].

Preventing adversarial disruptions in remote sensing classification requires models to adapt across varying spectral domains, withstand pixel-level noise, and maintain semantic coherence [7]. A robust defense must operate across spectral bands and spatial contexts, ensuring that adversarial influence does not propagate through learned feature hierarchies. Defensive strategies must also consider temporal and multi-sensor consistency to counter cross-domain adversarial transfer [8].

The resilience of remote sensing classification depends not only on the depth of the network but also on the inclusion of spectral attention, spatial redundancy, and geometric consistency within learning mechanisms [9]. By integrating spectral-aware defenses, classifiers can suppress the influence of adversarial signals while enhancing true class representations. Addressing the challenge of adversarial vulnerability in remote sensing, therefore necessitates defense architectures that go beyond traditional training, embedding contextual awareness and feature-level resistance into each phase of the classification pipeline [10].

1.1 Problem Statement

Remote sensing image classification systems are vulnerable to adversarial attacks that introduce perturbations across spectral dimensions, resulting in misclassification of land cover or environmental features. These perturbations exploit the sensitivity of deep learning models to minute spectral variations, particularly in high-dimensional data where spectral redundancy can be used against the model. In scenarios involving cross-sensor or multi-spectral satellite data, perturbations often shift class probabilities away from true labels, even under minor input distortion. Existing models lack mechanisms to verify spectral consistency across layers, leading to cascading feature misalignment under adversarial influence. The absence of cross-spectral attention makes it difficult to suppress unnatural variations injected by adversarial inputs. Most current defense methods either degrade clean image performance or lack adaptability across variable spectral bands. A defense architecture that can incorporate spectral awareness while preserving classification integrity is essential. Addressing this challenge requires a framework capable of reinforcing spectral fusion mechanisms, attending to relevant wavelength ranges, and filtering out perturbation-induced noise while maintaining spatial semantics and inter-band consistency. Enhancing adversarial robustness through spectral fusion is necessary to ensure trustworthy classification outcomes in diverse remote sensing tasks, including land mapping, environmental monitoring, and disaster response.

1.2. Motivation

Adversarial attacks in multi-spectral remote sensing challenge model integrity by disrupting inter-band feature alignment. Robustness in such environments depends on the model's ability to differentiate between spectral distortions and class-relevant patterns. Strengthening spectral attention during classification not only improves performance under normal conditions but also provides resilience against input manipulation. Integrating spectral fusion across attention layers enables more stable representations that are less influenced by perturbations. Motivated by the rising demand for accurate geospatial classification under real-world conditions, an effective spectral-aware defense strategy can significantly reduce the risk of misinterpretation and improve trust in remote sensing analytics.

1.3. Objective

The main objective is to design a defense-oriented classification framework that improves robustness against adversarial spectral perturbations without compromising clean sample accuracy. The approach aims to enhance spectral feature learning using an attention mechanism that aligns critical bands across the spectral domain. By integrating a cross-spectral attention mechanism into the classification pipeline, the goal is to enable models to selectively prioritize consistent spectral features while suppressing adversarial distortions. A robust fusion strategy will be incorporated to enforce redundancy elimination and maintain inter-band coherence under adversarial conditions. The objective also includes evaluating the proposed approach across multi-spectral and hyperspectral datasets to assess its adaptability, precision, and generalization capabilities. The objective encompasses the integration of spectral filtering with global spatial attention to improve the model's discrimination of legitimate versus adversarial feature patterns, ensuring resilience in large-scale remote sensing applications.

2. LITERATURE REVIEW

"Lean Attention VGG" [11] Embedded a compact convolutional pipeline by restructuring VGG with reduced parameters and depthwise separable layers. Spatial attention filters captured fine-grained landscape features using channel-weighted saliency maps. Lightweight modules preserved feature integrity while reducing computation. Data augmentation techniques enhanced robustness in various terrain textures.

"Segment-Adapt SAM Layers" [12] Employed a multi-level adaptation of the Segment Anything Model by redesigning mask propagation layers across shallow, mid, and deep features. Segment masks were iteratively refined using local-level guidance, followed by global context rectification. Patch-wise adaptations allowed flexible focus on irregular shapes in urban, vegetation, and aquatic zones. "Tea Drought Classifier" [13] Merged thermal, spectral, and SAR signals using a two-stream fusion block tuned for vegetation stress zones. Normalized Drought Index layers were computed through wavelet-based moisture transformation across channels. Gradient harmonization adjusted sensor inconsistencies using segment-focused weight realignment.

"Volcano Fusion Net" [14] crafted a dual-branch pipeline merging spectral-shape and textural-spatial cues from multi-band composites. High-dimensional convolutional patches were passed through a guided aggregation module that realigned irrelevant pixels using shape-prior enhancement. Feature fusion took place across hierarchical depth levels, supported by a decoder that filtered transitional textures. "CropText Temporal Fusion" [15] integrated satellite-derived time-series data with geo-tagged crop reports using a dual-attention encoder. Spatio-temporal layers handled seasonal variation and cloud occlusion using Gated Recurrent Maps. Text-based class anchors were embedded into a contrastive module that helped disambiguate overlapping crops. "VAN Ship Classifier" [16] customized the Vision Attention Network by integrating multi-head shape-aware tokens into the encoding layers. Positional awareness was fine-tuned for maritime texture variance using normalized anchor shifting. A ship-type-specific decoder was applied to extract discriminative traits across vessel subclasses. Cross-scale feature blending and edge-prior weighting enhanced contour preservation. "ResViT Attack Injector" [17] combined ResNet modules with transformer-based patches to generate feature-consistent adversarial triggers. Gradient-guided residual links injected localized perturbations at shallow layers. Transformer attention matrices were adjusted dynamically based on adversarial saliency maps. Patch-wise embeddings were selectively distorted using norm-bounded constraints.

"Privacy Adversarial Shield" [18] Structured an image sanitization method by generating adversarial examples to suppress identity traceability. Perturbations were crafted using targeted gradient masking while retaining original utility. A feature-steering module guided manipulations towards identity-critical zones, balancing privacy and integrity. "SAR Synthetic Gap Probe" [19] Deployed a benchmarking strategy comparing adversarial susceptibility between synthetic and real SAR samples in Automatic Target Recognition (ATR). Adversarial triggers were generated using projected gradient descent under controlled radar noise. A perturbation sensitivity matrix was computed for both domains. Robustness differences were analyzed using spectral variation heatmaps. "SmartGrid Threat Simulator" [20] Constructed a multi-agent simulation pipeline representing layered cyber-physical attacks in power grids. Each stage mimicked sensor jamming, communication spoofing, and energy rerouting. Defense layers included anomaly scoring, state prediction error correction, and real-time feedback loops.

"Fusion Attack Transfer" [21] designed a multi-step attack pipeline targeting joint systems combining image fusion and classification. Attack gradients were backpropagated across both fusion and classifier modules using shared loss paths. Perturbation consistency was ensured by aligning feature representations pre- and post-fusion. Targeted regions were identified through importance scoring across fused modalities. "Weighted Spectral-Spatial Shield" [22]combined transformer-driven spectral attention with spatial self-attention through a weight-balanced fusion layer. Spectral encoders captured wavelength-invariant patterns, while spatial encoders addressed geometric distortion. Fusion weights were optimized dynamically to suppress adversarial gradient paths. A patch-based consistency block stabilized token-level predictions. "AutoSafety Threat Review" [23] presented a structured examination of adversarial vulnerabilities across sensor, image, and control models in self-driving systems. Attack strategies included physical patch-based, digital perturbation, and spatio-temporal triggers. Defense mechanisms ranged from adversarial training and sensor redundancy to dynamic model switching. "Shallow Resilience Probe" [24] analyzed shallow models like decision trees and SVMs under adversarial pressure. Attack simulations employed gradient-free optimization and surrogate mapping. Model robustness was examined using perturbation thresholds and confidence shifts. Structural sparsity in shallow models limited the attack surface area. Adversarial transferability was assessed using common datasets. "Memory Malware Detector" [25] implemented memory snapshot scanning via forensics tools feeding into a convolutional neural

classifier. Byte-level sequences were encoded as grayscale images, preserving access patterns. Feature extraction relied on a multi-kernel CNN configured for sparse detection zones. Detection accuracy improved through memory-section-specific normalization.

"MSRF" [26] Constructed a pixel-level image fusion model aligning multi-spectral sources using directional gradient statistics. A moment-matching module preserved edge sharpness by balancing intensity transitions between channels. High-frequency details were retained through gradient field compensation, while contrast enhancement used local moment weights. Fusion maps were generated by adapting channel dominance per region. An edge-preserving regularizer maintained object boundaries across varying resolutions. "A3OD" [27] Deployed adversarial triggers to manipulate object detection pipelines trained on remote sensing datasets. Perturbations were crafted using gradient-based attacks targeting detection heads and region proposal networks. Patch-wise perturbations disrupted bounding box confidence, misaligning spatial anchors.

Bio-inspired optimization techniques have been effectively applied across various domains for enhancing robustness and feature learning [28]-[42]. In remote sensing adversarial defense, such strategies can also be incorporated to fine-tune spectral fusion, attention weighting, or adversarial training schedules for improved adaptability.

3. CROSS-SPECTRAL ATTENTION-FUSED VISION TRANSFORMER (CSAF-ViT)

CSAF-ViT is designed for robust remote sensing image classification. It integrates spectral-aware attention, patch-based encoding, and adversarial training to enhance feature fusion, spatial reasoning, and resilience against perturbations, enabling accurate classification under complex spectral conditions and hostile input manipulations.

3.1. Spectral Band Decomposition

Remote sensing imagery acquired from satellite or airborne platforms contains multiple spectral channels representing surface reflectance across different electromagnetic regions. These channels are tightly coupled with land cover properties, atmospheric conditions, and object materials. To initiate the CSAF-ViT pipeline, the first computational objective is to isolate these spectral components for targeted modelling. The process begins with organizing the input image tensor $I \in R^{H \times W \times C}$, where H and W represent the spatial dimensions and C denotes the total number of spectral bands.

Each spectral slice from the input tensor is extracted to form an individual spectral representation. Let $I_c \in R^{H \times W}$ denote the image representation from the c-th spectral band.

$$I_c = I[:, :, c] \forall c \in \{1,2, \ldots., C\} \quad (1)$$

This operation forms a stack of C mono-spectral images, which are further subjected to controlled normalization processes for spectral alignment. The pixel intensities of each band are rescaled to ensure compatibility in subsequent feature encoding stages. The normalization transformation is represented as follows

$$\tilde{I}_c(i,j) = \frac{I_c(i,j) - \mu_c}{\sigma_c} \quad (2)$$

where μ_c and σ_c denote the mean and standard deviation of the band c, computed independently to retain the natural distribution of spectral intensity. This operation yields standardized spectral slices $\tilde{I}_c$ capable of preserving intrinsic radiometric patterns critical for downstream spectral attention.

To reduce noise variance and enforce structure within each spectral channel, a low-pass Gaussian smoothing operation is applied over each normalized slice.

$$\bar{I}_c = G_\sigma * \tilde{I}_c \quad (3)$$

Here, G_σ represents a Gaussian kernel with a defined standard deviation σ, and $*$ denotes the convolution operator. The outcome, $\bar{I}_c$, highlights prominent spatial-spectral features by attenuating abrupt pixel-level fluctuations, which are often susceptible to adversarial alterations.

Once denoised, spectral slices are reorganized into a list of tensors suitable for parallel processing. For each spectral band c, the filtered matrix $\bar{I}_c$ is reshaped and encoded into a 3D tensor of shape $1 \times H \times W$, yielding a batch-aligned spectral tensor $S \in R^{C \times H \times W}$

$$S[c, :, :] = \bar{I}_c \quad (4)$$

This tensorized representation supports batch-wise attention operations, facilitating spatial-spectral

interactions in the next stage. The channel-aligned format also supports parallelization on GPU memory architectures, significantly accelerating spectral encoding in deep pipelines.

Each spectral band in S is further assessed for structural variability to prioritize regions carrying discriminative spectral information. This is realized by computing the spectral entropy E_c for each band as

$$E_c = -\sum_{k=1}^{L} p_k^c \cdot log(p_k^c) \quad (5)$$

Where p_k^c denotes the probability of the intensity value k in band c, and L represents the total number of discrete intensity levels. The entropy measure determines the complexity and diversity within the band, guiding downstream spectral attention layers to weigh high-entropy channels more strongly in fusion.

After entropy evaluation, the full spectral set S is packaged with corresponding statistical descriptors, bandwise entropy, mean, and variance as auxiliary signals for the cross-spectral attention stage. This metadata supports selective weighting in fusion, assisting in countering adversarial perturbations that tend to exploit low-entropy or homogenous spectral responses.

The decomposition phase concludes with a consistent tensor format ready for spectral-wise feature encoding, preserving both raw intensity details and structured statistical characteristics. This representation becomes foundational for extracting cross-band interactions, which are essential for the adversarially robust learning objective in CSAF-ViT. Each spectral channel maintains its native integrity while aligning structurally with others, enabling meaningful correlation analysis under hostile pixel disturbances. The formatted tensor also ensures that subsequent cross-spectral attention operates on clean, normalized, and entropy-aware representations, reinforcing the overall resilience of the classification model in operational deployments.

3.2. Band-wise Feature Encoding

Each filtered and normalized spectral band obtained from the spectral decomposition process carries radiometric characteristics specific to its wavelength sensitivity. These characteristics must be transformed into a compact and learnable feature space while preserving spectral discriminability and robustness against adversarial perturbations. To achieve this, a dedicated encoder operates independently on every channel within the spectral tensor $S \in R^{C \times H \times W}$, where C indicates the number of spectral bands, and H, W represent spatial dimensions.

For each spectral slice $\bar{I}_c \in R^{H \times W}$, a lightweight convolutional encoder $\Phi_c(\cdot)$ is employed to extract hierarchical features. The initial transformation applies a convolutional operation to emphasize local textural structures:

$$F_c^{(1)} = \delta\left(BN\left(W_c^{(1)} * \bar{I}_c + b_c^{(1)}\right)\right) \quad (6)$$

Where $W_c^{(1)}$ and $b_c^{(1)}$ denote the kernel weights and biases for the first convolutional layer in the encoder for the band c, $BN(\cdot)$ applies batch normalization to stabilize gradients, and $\delta(\cdot)$ is a ReLU activation function. This feature map $F_c^{(1)} \in R^{H \times W \times d_1}$ reflects initial edge and texture patterns embedded within each band.

To increase the receptive field and compress redundant details, a strided convolution is introduced in the next stage of the encoder

$$F_c^{(2)} = \delta\left(BN\left(W_c^{(2)} * F_c^{(1)} + b_c^{(2)}\right)\right) \quad (7)$$

This transformation downsamples the spatial resolution while preserving informative structures. The output feature map $F_c^{(2)} \in R^{H/2 \times W/2 \times d_2}$ increases spectral compactness, which is critical for memory-efficient fusion in later attention blocks.

To introduce non-local context without expanding the network depth, a depthwise separable convolution layer is employed, decomposing standard convolution into depthwise and pointwise operations. This operation enhances channel-specific learning

$$F_c^{(3)} = \delta\left(BN\left(W_c^{(3,d)} \odot F_c^{(2)} + W_c^{(3,p)} *\cdot + b_c^{(3)}\right)\right) \quad (8)$$

Here, $W_c^{(3,d)}$ and $W_c^{(3,p)}$ are the depthwise and pointwise kernels, $\odot$ represents depthwise convolution across spatial axes, and $*$ signifies pointwise mixing across channels. This architectural choice ensures high efficiency while retaining expressive power.

In order to maintain consistency in feature scales across all spectral bands, a spectral alignment

transformation is applied. This transformation reprojects each feature tensor to a unified latent dimension d using a learnable projection matrix P_c

$$Z_c = P_c \cdot Flatten\left(F_c^{(3)}\right) \quad (9)$$

where $Z_c \in R^{(H/2 \cdot W/2) \times d}$ represents the encoded token sequence for the spectral band c, and $Flatten(\cdot)$ reshapes the feature map into a 2D token representation suitable for cross-spectral attention computation.

To enforce spectral sparsity and suppress adversarially induced noise patterns, a channel-wise dropout mask $M_c \in \{0,1\}^d$ is applied stochastically over the latent space

$$Z_c' = M_c \odot Z_c \quad (10)$$

The dropout mask M_c randomly turns off feature dimensions during training, promoting spectral generalization and resilience. Each encoded tensor Z_c' is spectrally conditioned, entropy-aware, and free of low-level redundancies.

Encoded tensors across all C bands are collated to form a unified encoding bank

$$Z = [Z_1', Z_2', \ldots, Z_C'] \quad (11)$$

This bank $Z \in R^{C \times (H/2 \cdot W/2) \times d}$ becomes the fundamental structure for computing cross-spectral attention maps. All tensors in this representation space are now harmonized across spatial scale, spectral variance, and latent dimensionality, making them suitable for integration under shared attention heads.

Each band-wise encoder functions independently without parameter sharing to preserve spectral uniqueness. At the same time, architectural symmetry across encoders ensures computational uniformity and scalability. The encoding step ensures that raw spectral slices are converted into compressed, noise-filtered, and semantically rich representations, which can later be compared, fused, and reweighted through attention operations in the next cross-spectral fusion stage. The resulting tensors serve as robust descriptors of both structural geometry and spectral contrast, enabling stronger resilience under adversarial input scenarios targeting spectral dependencies.

3.3. Query-Key-Value Generation for Attention

Encoded spectral tokens, structured as latent tensors from the band-wise encoding process, form the foundation for attention-driven feature fusion. Each spectral embedding $Z_c' \in R^{N \times d}$, where $N = H/2 \cdot W/2$ and d denotes the latent dimension, holds localized structural and spectral semantics. To compute inter-band relationships and calibrate spectral dependencies, these embeddings are linearly transformed into three separate representations: queries, keys, and values. These vectors define the interaction dynamics among the spectral embeddings through scaled attention mechanisms.

Each spectral token set Z_c' undergoes three distinct linear transformations using separate projection matrices to yield the respective tensors

$$Q_c = Z_c' W_c^Q, \quad K_c = Z_c' W_c^K, \quad V_c = Z_c' W_c^V \quad (12)$$

Here, $Q_c, K_c, V_c \in R^{N \times d_a}$, with d_a representing the attention-specific dimensionality. The matrices $W_c^Q, W_c^K, W_c^V \in R^{d \times d_a}$ are learnable parameters associated with the spectral band c. These transformations allow spectral descriptors to be projected into an attention-interpretable space, with queries driving the information extraction, keys offering matching signals, and values representing feature payloads.

Once projected, spectral embeddings across bands are aligned by stacking the transformed tokens into multi-band matrices. The collective matrices are defined as

$$Q = [Q_1; Q_2; \ldots; Q_c], \; K = [K_1; K_2; \ldots; K_c], \; V = [V_1; V_2; \ldots; V_c] \quad (13)$$

Each matrix $Q, K, V \in R^{(C \cdot N) \times d_a}$ aggregates all tokenized spectral representations into unified spaces. The ordering ensures that tokens are indexed by spectral origin, preserving traceability throughout attention computation. This alignment is essential for evaluating spectral-level relevance across bands and guiding value integration in upcoming stages.

Before performing attention-weight calculations, spectral relevance is modulated using entropy-informed scaling to emphasize spectrally discriminative tokens. Let λ_c denote the normalized entropy score for band c, computed in the

decomposition phase. Each band-specific query vector is scaled using this factor:

$$\tilde{Q}_c = \lambda_c \cdot Q_c \quad (14)$$

This entropy-aware modulation introduces attention asymmetry favouring high-variability bands, improving robustness against adversarial attacks that exploit low-entropy regions. All scaled query tensors are then assembled into the global query matrix $\tilde{Q} \in R^{(C\cdot N)\times d_a}$, preserving the original token order.

A pairwise attention interaction is computed by measuring the scaled dot product between the global query and key matrices, normalized by the square root of the dimensionality d_a

$$A = Softmax\left(\frac{\tilde{Q}\cdot K^T}{\sqrt{d_a}}\right) \quad (15)$$

The matrix $A \in R^{(C\cdot N)\times(C\cdot N)}$ encodes all spectral-token affinities across bands. Each row corresponds to a query token's alignment scores with every key token. The softmax function ensures probabilistic interpretation, allowing spectral tokens to attend to semantically compatible counterparts from other bands selectively.

To refine attention gradients and reduce susceptibility to spurious correlations, a spectral regularization mask $R \in R^{(C\cdot N)\times(C\cdot N)}$ is applied to the attention matrix

$$A_{reg} = A \odot R \quad (16)$$

The regularization mask R imposes structural constraints by penalizing spatially distant or spectrally irrelevant token interactions. This guides the model to focus on physically and spectrally plausible correspondences. The operation $\odot$ indicates element-wise multiplication, maintaining differentiability for gradient-based optimization.

To preserve inter-band relationships during backpropagation, each attention vector is constrained through adaptive normalization. An attention-balancing function ensures uniform attention spread over valid spectral regions:

$$A_{norm}(i,:) = \frac{A_{reg}(i,:)}{\sum_j A_{reg}(i,j) + \epsilon} \quad (17)$$

The inclusion of a small constant ϵ prevents division instability. This normalization facilitates robust training dynamics and avoids overfitting to dominant bands. The normalized attention matrix A_{norm} is passed to the fusion layer, where it will be used to compute context-aware spectral representations.

All steps involved in query-key-value projection are differentiable and integrated seamlessly with the encoding outputs from the previous stage. The process ensures spectral relevance is captured not through static averaging, but through dynamic relationships driven by entropy-conditioned interaction matrices. This structure enables meaningful, controllable attention flow that can prioritize critical spatial-spectral cues even under adversarial deformation of input signals. Each projected embedding maintains position and spectral origin while transforming into a learned latent space. The output matrices Q, K, V form the core of the CSA block, ensuring cross-spectral relationships are contextually grounded. These representations are the interface between raw spectral content and the fusion mechanism that forms the backbone of adversarially robust classification. The structure constructed here becomes a spectral-symmetric attention scaffold that will next guide selective value weighting across all spectral interactions.

3.4. Cross-Spectral Attention Map Computation

The structured triplet embeddings Q, K, V generated across all spectral channels, provides the necessary foundation for learning spectral dependencies in a controlled and responsive manner. The goal in this phase is to compute cross-spectral attention maps that allow contextually relevant features from one spectral band to guide the interpretation of another, thereby mitigating vulnerability to pixel-level perturbations and boosting semantic alignment across bands.

Let $Q \in R^{T\times d_a}$ and $k \in R^{T\times d_a}$ represent the stacked query and key matrices, respectively, where $T = C \cdot N, C$ is the number of bands, and N is the number of spatial tokens per band. The attention logits are computed using the scaled dot product

$$M = \frac{Q\cdot K^T}{\sqrt{d_a}} \quad (18)$$

This raw attention matrix $M \in R^{T\times T}$ encodes the unnormalized spectral affinity between every

query-key pair across all bands. Each entry $M_{i,j}$ measures how strongly the token i from one spectral channel aligns with a token j from potentially another, forming a rich matrix of inter-band semantic relationships.

To obtain valid probabilities and ensure focus on relevant associations, a softmax transformation is applied to normalize the logits along each row

$$A_{raw}(i,:) = Softmax(M(i,:)) \quad (19)$$

The resulting matrix $A_{raw} \in R^{T\times T}$ holds attention weights that sum to one across each row, preserving interpretability and numerical stability. These raw attention values, though normalized, may still suffer from diffuse focus or irrelevant long-range attention in high-dimensional remote sensing data.

To address this issue and embed spectral priors into the attention mechanism, a cross-band distance mask $D \in R^{C\times C}$ is constructed, where each entry $D_{\mu,\eta}$ captures the relative band-level dissimilarity derived from entropy or statistical divergence. This matrix is expanded spatially to match the token dimensions:

$$D^*(i,j) = D_{c_i,c_j} \text{where}$$
$$c_i = \lfloor i/N \rfloor, c_j = \lfloor j/N \rfloor \quad (20)$$

The extended distance-aware penalty matrix $D^* \in R^{T\times T}$ is applied over the attention map to downscale semantically distant or statistically uncorrelated spectral relationships. This is achieved through exponential decay-based weighting:

$$A_{masked}(i,j) = A_{raw}(i,j) \cdot exp(-\beta \cdot D^*(i,j)) \quad (21)$$

here, β is a tunable decay factor controlling the sharpness of the spectral penalty. The mask adaptively reduces irrelevant cross-band influences while preserving proper transitions. After this modulation, a row-wise renormalization is necessary to restore the probabilistic structure:

$$A(i,:) = \frac{A_{masked}(i,:)}{\sum_{k=1}^{T} A_{masked}(i,k) + \epsilon} \quad (22)$$

The constant ε ensures stability in edge cases involving zero summations. The resulting matrix $A \in R^{T\times T}$ is the final cross-spectral attention map containing adaptively regulated weights connecting every spatial token in every spectral band to all others.

To facilitate interpretability and attention traceability, an optional residual emphasis term is incorporated, where spectral self-attention is boosted by a diagonal dominance term $\alpha \in [0,1]$

$$A_{final}(i,j) = \begin{cases} (1-\alpha)\cdot A(i,j) + \alpha, & if\ i = j \\ (1-\alpha)\cdot A(i,j), & otherwise \end{cases} \quad (23)$$

This formulation increases the importance of self-token interactions, enhancing feature stability for isolated or rare patterns, especially under adversarial interference. The final matrix A_{final} serves as the gating mechanism in the fusion process, controlling how much information flows from every spectral token to every other.

The attention map computation strategy detailed here offers more than conventional soft alignment. The entire structure is entropy-regularised, spectrally informed, spatially balanced, and adversarially resilient. Spectral fusion is controlled through A_{final} is capable of emphasizing regions where meaningful interactions occur across bands, such as edges, thermal anomalies, or vegetation structures, while suppressing noise-prone, homogenous areas.

All operations in this step maintain differentiability, allowing smooth backwards propagation during training. The attention gradients help refine spectral keys and queries in earlier stages, contributing to a tighter, more specialized representation per band. This ensures that final fused features retain physically grounded, semantically rich correlations derived from structurally controlled attention weights. The output of this step feeds directly into value-weighting modules, forming the next critical step in cross-spectral aggregation.

3.5. Spectral Feature Reweighting

The cross-spectral attention matrix $A_{final} \in R^{T\times T}$, constructed through entropy-aware and spectrally regularized mechanisms, forms the core of contextual information flow between spectral tokens. Each row in this matrix determines how a given token integrates information from all other tokens distributed across spectral bands. To initiate spectral feature reweighting, the attention-guided transformation of the value matrix

$V \in R^{T \times d_a}$ is executed through weighted aggregation.

The reweighted representation for each query token is computed by performing a dense matrix multiplication between the final attention matrix and the stacked value matrix

$$F = A_{final} \cdot V \quad (24)$$

The resulting tensor $F \in R^{T \times d_a}$ holds the fused spectral-spatial descriptors, where each row corresponds to a token updated through weighted aggregation of inter-band contextual features. This step facilitates multispectral information fusion by allowing tokens to borrow semantically relevant content from neighboring or distant spectral sources based on the computed attention distribution.

To enhance representational clarity and enforce spectral disambiguation, a spectral context modulation function is introduced. This modulation acts as a channel-wise gate to selectively enhance or suppress token-wise activations, based on the learned global statistics:

$$G_c = \sigma(W_g \cdot Pool(F_c) + b_g) \quad (25)$$

Here, $F_c \in R^{N \times d_a}$ refers to the token set originating from the spectral band c, extracted from F. The function $Pool(\cdot)$ computes a global descriptor using average pooling, $W_g \in R^{d_a \times d_a}$ is a learnable weight matrix, b_g is the bias vector, and $\sigma(\cdot)$ denotes a sigmoid function. The modulation gate $G_c \in R^{1 \times d_a}$ dynamically calibrates channel importance across spectral contexts.

The modulated feature tensor $\hat{F}_c \in R^{N \times d_a}$ for each spectral group, it is obtained through channel-wise multiplication.

$$\hat{F}_c = F_c \odot G_c \quad (26)$$

The operator $\odot$ signifies broadcasted element-wise multiplication. The outcome retains the original spatial structure while incorporating spectral sensitivity derived from attention-driven global descriptors. Each token embedding now encapsulates both fine-grained and globally modulated spectral attributes, increasing robustness against irregular or adversarial patterns.

To restore positional alignment and prepare for patch generation, the reweighted and modulated tokens $\hat{F}_c$ are reshaped back to their spatial layout. This reverse transformation maps the sequence of tokens into 2D grid structures aligned with the downsampled spatial dimensions $H/2 \times W/2$

$$\hat{M}_c - Reshape(\hat{F}_c) \quad (27)$$

The resulting tensor $\hat{M}_c \in R^{H/2 \times W/2 \times d_a}$ represents the fused spectral-spatial feature map for spectral band c, now enriched through global attention flow and spectral context refinement.

To ensure compatibility and uniformity for later concatenation across bands, all such feature maps are aligned along the channel axis. This is done using a spectral unification block that linearly projects each band's reweighted map into a shared dimensional space d_u. The projection is defined as:

$$u_c - \hat{M}_c \cdot W_u^c \quad (28)$$

$W_u^c \in R^{d_a \times d_u}$ denotes the band-specific projection matrix, and $u_c \in R^{H/2 \times W/2 \times d_u}$ becomes the unified reweighted map for the corresponding spectral channel. This transformation guarantees channel-wise comparability and supports seamless stacking for transformer-based modelling in subsequent steps.

The complete reweighted tensor set is now consolidated across bands:

$$U = [u_1, u_2, \ldots, u_c] \quad (29)$$

The stacked tensor $U \in R^{C \times H/2 \times W/2 \times d_u}$ encodes reweighted spectral features from all bands, ensuring that each token has absorbed critical inter-spectral cues. This tensor structure forms the pre-fusion representation carrying spatial locality, spectral awareness, and attention-mediated reweighting necessary for effective downstream patch embedding and positional alignment.

The feature reweighting mechanism built on attention-driven value transformation and context-guided channel modulation facilitates selective signal enhancement. It removes the burden of treating all spectral inputs equally, instead emphasizing bands with meaningful

contrast and texture for classification. This strategy prevents noise propagation and information dilution, strengthening the learning path against spectral adversaries attempting to distort feature hierarchies. The reweighted feature structure proceeds directly into the fusion layer, where spatial patching and transformer embedding are executed to continue the robustness-aware classification path.

3.6. Feature Fusion Across Spectral Bands

The reweighted spectral descriptors $\in R^{C\times H/2\times W/2\times d_u}$, obtained after spectral attention modulation and context gating, serve as structured blocks of informative content representing localized patterns and their corresponding spectral dependencies. To construct a unified representation that encapsulates both spectral diversity and spatial granularity, feature fusion is initiated across all C spectral dimensions at each spatial location. The fusion is performed to maximize semantic consistency and suppress redundancy introduced by repetitive spectral overlaps.

At each spatial index (i,j), a feature vector $V_{i,j} \in R^{C\times d_u}$ is extracted from the stacked tensor U, defined as

$$V_{i,j} = [u_1(i,j,:), u_2(i,j,:), \dots, u_C(i,j,:)] \quad (30)$$

This vector captures all spectral channel activations centred on a particular spatial coordinate. A dynamic fusion gate $\Gamma_{i,j} \in R^C$ is generated using a shared soft attention unit to regulate the contribution of each spectral component during fusion. The attention scores are derived by compressing each spectral vector into a scalar weight through

$$\Gamma_{i,j}(c) = \frac{exp(\phi^\top \cdot u_c(i,j,:))}{\sum_{k=1}^{C} exp(\phi^\top \cdot u_k(i,j,:))} \quad (31)$$

Here, $\phi \in R^{d_u}$ is a learnable parameter vector, enabling the gating to be sensitive to the spectral vector's relevance in the latent space. The attention values in $\Gamma_{i,j}$ serve as a normalized weighting function, assigning more emphasis to discriminative spectral bands and attenuating noisy or redundant ones.

Using the computed fusion gate, the final fused vector at the coordinate (i,j), denoted by $F_{i,j} \in R^{d_u}$, is formulated as a weighted summation over spectral axes

$$F_{i,j} = \sum_{c=1}^{C} \Gamma_{i,j}(c) \cdot u_c(i,j,:) \quad (32)$$

This operation is repeated across the entire spatial grid to form a unified feature map $\in R^{H/2\times W/2\times d_u}$. The resulting tensor integrates spectral diversity into a single channel at each pixel while preserving spatial context, delivering a spectrally informed representation free from redundant alignment conflicts or positional mismatches.

To further enforce inter-spectral consistency and suppress feature drifts introduced by uneven channel scaling, a spectral consistency regularizer is imposed. A spectral correlation matrix $\Psi \in R^{C\times C}$ is computed using dot-product similarities among channel-wise features

$$\Psi(p,q) = \frac{\sum_{i,j} u_p(i,j,:) \cdot u_q(i,j,:)}{\sqrt{\sum_{i,j} \|u_p(i,j,:)\|_2^2 \cdot \sum_{i,j} \|u_q(i,j,:)\|_2^2}} \quad (33)$$

This matrix assesses how strongly two spectral channels correlate over the spatial domain. A regularization loss is attached to minimize the high correlation between irrelevant bands and maximize diversity in relevant ones. The spectral orthogonality loss L_{so} is defined as

$$L_{so} = \sum_{p\neq q} (\Psi(p,q))^2 \quad (34)$$

The minimization of L_{so} drives orthogonality between channels, ensuring that the fusion gate operates on decoupled and informative inputs. This regularization also contributes to robustness, as it prevents the model from collapsing into trivial representations vulnerable to perturbations across similar bands.

After fusion, the global feature tensor $F \in R^{H/2\times W/2\times d_u}$ is optionally passed through a refinement block, which utilizes a residual bottleneck architecture to sharpen activations and suppress aliasing. This is achieved through a pair of convolutional layers

$$F' = \delta\left(BN\left(W_2 * \delta(BN(W_1 * F))\right)\right) + F \quad (35)$$

Here, $W_1, W_2 \in R^{3\times 3\times d_u}$ represent convolution filters, $BN(\cdot)$ indicates batch normalization, and

$\delta(\cdot)$ denotes a ReLU activation. The skip connection ensures gradient stability and preserves low-frequency signals.

The fused and refined feature map now exhibits spatial integrity and spectral alignment, ready for partitioning into fixed-size embeddings for transformer processing. This feature tensor serves as the spatially aware and spectrally condensed carrier of all meaningful semantics extracted from the input hyperspectral or multispectral imagery, bridging the gap between localized perception and transformer-based reasoning in the upcoming stages.

3.7. Patch Generation from Fused Map

The spectrally integrated and spatially refined representation $F' \in R^{H/2 \times W/2 \times d_u}$ encapsulates dense semantic patterns that combine structural information and adversarially calibrated spectral attention. In order to interface this unified feature map with the transformer framework, a transformation into fixed-size non-overlapping patches is performed. This patchification process must preserve locality, spectral alignment, and boundary integrity, enabling transformer attention heads to operate meaningfully on discrete visual regions.

Let $P \times P$ denote the patch resolution, where P is the predefined patch size. The total number of horizontal and vertical patches is given by $N_h = \frac{H/2}{p}$ and $N_w = \frac{W/2}{p}$, assuming P divides the spatial dimensions exactly. The fused feature map is divided into spatially continuous blocks, each block $P_{i,j} \in R^{P \times P \times d_u}$ corresponding to a unique spatial region. This can be formally expressed as

$$P_{i,j} = F'[i \cdot P:(i+1) \cdot P, j \cdot P:(j+1) \cdot P, :] \quad (36)$$

where $i \in \{0,1,\ldots,N_{h-1}\}$ and $j \in \{0,1,\ldots,N_{h-1}\}$. Each $P_{i,j}$ captures a local subregion, preserving the spectral fusion and attention semantics obtained earlier. These blocks serve as the atomic inputs for token generation in the transformer.

Every spatial patch is transformed into a flat vector by reshaping its $P \times P \times d_u$ structure into a one-dimensional embedding of length $P^2 \cdot d_u$. This transformation is denoted as

$$z_{i,j} = Flatten(P_{i,j}) \in R^{P^2 \cdot d_u} \quad (37)$$

The result is a sequence of patch vectors $z_{i,j}$, which represent discrete visual units with embedded spectral correlations. To unify these vectors for transformer input, a linear projection is applied to each flattened vector to map it to a fixed-dimensional latent space R^{d_p}, ensuring consistency with the transformer model

$$e_{i,j} = W_p \cdot z_{i,j} + b_p \quad (38)$$

where, $W_p \in R^{d_p \times (P^2 \cdot d_u)}$ and $b_p \in R^{d_p}$ are learnable parameters used for embedding generation. The transformation $e_{i,j} \in R^{d_p}$ encodes local spatial features as well as spectral abstraction in a transformer-compatible token format.

All patch embeddings $e_{i,j}$ are collected to form the patch token sequence $\varepsilon \in R^{N \times d_p}$, where $N = N_h . N_w$ is the total number of patches. This sequence reflects the flattened spatial structure of the image, retaining localized descriptors in a format suitable for global attention operations. To maintain spatial coherence and enable positional context modelling, positional encoding is incorporated at this stage.

The positional embedding matrix $p_{pos} \in R^{N \times d_p}$ is constructed using learnable sinusoidal or parameterized vectors, and added element-wise to the patch embeddings

$$\tilde{\varepsilon} = \varepsilon + p_{pos} \quad (39)$$

The combined embedding $\tilde{\varepsilon}$ carries spatial-awareness, allowing transformer heads to distinguish between spatially adjacent and distant patches. The structure formed at this stage is transformer-ready, enabling the model to operate over a tokenized representation where each token holds compact spectral-spatial knowledge of its respective region.

An optional classification token $e_{cls} \in R^{d_p}$ is prepended to the sequence, serving as a global aggregator for downstream class prediction. The final input to the transformer encoder is constructed as

$$e_{input} = [e_{cls}; \tilde{e}_1; \tilde{e}_2; \ldots, \tilde{e}_N] \quad (40)$$

This structure $\varepsilon_{input} \subset R^{(N+1)\times d_p}$ conforms to the ViT framework's expected input format, maintaining a balance between localized patch information and global class-level abstraction.

To ensure stable training dynamics and preserve norm stability across patch tokens, layer normalization is optionally applied over the sequence

$$\hat{\varepsilon}_{input} = LayerNorm(\varepsilon_{input}) \quad (41)$$

The normalized token set $\hat{\varepsilon}_{input}$ guarantees uniform input distribution before transformer encoding, mitigating the risk of divergence due to patch-wise variance or class token saturation.

The patch generation step described above transforms a dense, spectrally rich fused image into a structured sequence of visual tokens embedded with spectral coherence, adversarial resilience, and local-global spatial alignment. Each token contributes a quantized perspective of the image content, empowering the transformer layers to reason over globally contextualised yet semantically diverse inputs, essential for robust classification in remote sensing environments.

3.8. Linear Embedding of Patches

The output of the patch generation process yields a structured sequence $\varepsilon_{input} \in R^{(N+1)\times d_p}$, where each patch embedding captures spectral-spatial coherence in a localized region, and an optional classification token is prepended for global aggregation. Before passing this sequence into transformer encoder layers, a dedicated linear embedding phase is applied to enhance alignment across token dimensions, minimize local variation, and optimize gradient flow for attention-based modelling.

Let each token in ε_{input} be denoted as $e_t \in R^{d_p}$, where $t \in \{0,1,\ldots,N\}$. To reinforce projection uniformity and capture linear abstractions across all tokens, a shared transformation is applied using a learnable projection matrix $W_e \in R^{d_{prof} \times d_p}$, where d_{proj} is the final transformer-compatible embedding dimension. Each token is projected as

$$T_t = W_e \cdot e_t + b_e \quad (42)$$

where, $b_e \in R^{d_{proj}}$ denotes a trainable bias term. The result $T_t \in R^{d_{proj}}$ becomes a refined token embedding where the full representation lies in a normalized and compact latent space. The set of all tokens forms the sequence $T \in R^{(N+1)\times d_{prof}}$, which serves as the working input for the transformer.

To minimize distortions induced by outliers or band-specific irregularities in token magnitude, a feature-wise standardization is introduced across the embedding dimension. The standardized token $\tilde{T}_t$ is computed by subtracting the mean and scaling by the standard deviation for each feature

$$\tilde{T}_t(k) = \frac{T_t(k) - \mu_k}{\sigma_k} \quad (43)$$

where $\mu_k$and σ_k represent the mean and standard deviation across tokens at the feature index k, computed over the batch. This normalization enhances the stability of the transformer's internal representations, ensuring that different attention heads do not bias toward specific tokens based on unbalanced magnitudes.

To preserve the hierarchical nature of patches and mitigate the risk of spectral locality collapse, a residual pathway is introduced. The original token e_t is passed through a skip connection, scaled by a learnable gating coefficient $\lambda \in [0,1]$

$$H_t = \lambda \cdot \tilde{T}_t + (1-\lambda) \cdot e_t \quad (44)$$

This residual fusion $H_t \in R^{d_{prof}}$ maintains alignment with the pre-projection structure while benefiting from the expressiveness of the projection space. The operation provides resilience against local distortions and adversarial manipulations by ensuring that the original spectral cues are not discarded during embedding.

To introduce redundancy-aware semantic abstraction and reduce token-to-token variation, a multi-head aggregation block is introduced, which processes the sequence $\{H_t\}$ through grouped linear attention filters. Let M represent the number of attention branches, each operating independently over the projected features. For each head m, an independent projection $W_m \in R^{d_{prof} \times d_{prof}/M}$ is applied

$$A_t^{(m)} = W_m \cdot H_t \quad (45)$$

The projected sub-vectors $A_t^{(m)} \in R^{d_{proj}/M}$ are concatenated across all heads to form a joint representation

$$A_t = Concat\left(A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(M)}\right) \quad (46)$$

This embedding $A_t \in R^{d_{proj}}$ possesses multi-view semantics drawn from diverse linear filters. The inclusion of head-wise projections ensures diverse encoding of the same token space, improving robustness and enhancing generalization under spectral or adversarial variation.

To enforce embedding-level consistency and compress high-dimensional variation into a bounded subspace, a nonlinear compression is applied using a gated feedforward unit. The final embedding $\varepsilon_t \in R^{d_{prof}}$ is given by

$$\varepsilon_t = \delta\left(W_g \cdot A_t + b_g\right) \quad (47)$$

In this expression, $W_g \in R^{d_{prof} \times d_{prof}}$, and $\delta(\cdot)$ is a GELU activation function. This transformation enables a smooth, bounded, and differentiable nonlinearity that adapts well to attention-driven modelling.

All token embeddings ε_t are assembled into a sequence matrix $\varepsilon_{seq} \subset R^{(N+1) \times d_{prof}}$, containing context-aware, spectrally normalized, and structurally preserved patch descriptors. Each vector reflects the cumulative effects of spatially aligned projection, residual memory preservation, multi-head abstraction, and nonlinear adaptation. This matrix is then used as the definitive input for the transformer encoder, ensuring high-fidelity modeling and adversarial resilience during remote sensing classification. The entire embedding strategy strengthens the positional and contextual reliability of patch tokens, aligns spectral cues with global attention capacity, and guarantees the downstream transformer operates on semantically potent and adversarially stable inputs. The linear embedding layer thus completes the reformatting of raw fused features into structured tokens ready for transformer-based spatial reasoning.

3.9. Positional Information Encoding

The sequence of token embeddings $\varepsilon_{seq} \in R^{(n+1) \times d_{proj}}$, generated through linear projection and multi-head abstraction, represents content-rich patch descriptors. These tokens, while semantically robust, lack explicit spatial structure since the transformer architecture operates without inherent locality awareness. To compensate for this structural deficiency, positional information encoding is introduced, embedding spatial cues directly into the token representations to guide attention-based reasoning.

Let the total number of tokens be $T = N + 1$, where N represents the number of patches, and the additional token corresponds to the classification embedding. Each token $\varepsilon_t \in R^{d_{prof}}$ must be augmented with spatial context, allowing the transformer to infer positional dependencies during global attention computation. This is achieved by designing a positional embedding matrix $P_{enc} \in R^{T \times d_{proj}}$, where each row $P_{enc}(t)$ encodes the position of token t.

To ensure smooth interpolation and gradient-friendly properties, sinusoidal functions are used for encoding the spatial index. The sinusoidal formulation for the i-th dimension of the t-th position vector is given as:

$$p_{enc}(t, 2i) = sin\left(\frac{t}{10000^{2i/d_{proj}}}\right) \quad (48)$$

$$p_{enc}(t, 2i+1) = cos\left(\frac{t}{10000^{2i/d_{proj}}}\right) \quad (49)$$

This encoding yields high-frequency components for small positional indices and low-frequency patterns for larger indices, allowing the model to distinguish relative and absolute positions efficiently across long sequences. The alternating sine and cosine patterns help capture both linear and cyclic patterns in token ordering, preserving structural transitions from top-left to bottom-right patch order.

Each token embedding ε_t is updated through a direct summation with its corresponding positional vector

$$H_t^{pos} = \varepsilon_t + p_{enc}(t) \quad (50)$$

The updated embedding $H_t^{pos} \in R^{d_{proj}}$ now encapsulates both semantic features and spatial identity. The composite sequence $H_t^{pos} \in R^{T \times d_{proj}}$ becomes a spatially aware representation, enabling the transformer to weigh tokens based not only on their content relevance but also their positional role in the image layout.

To enhance positional adaptiveness and learn image-dependent spatial hierarchies, a learnable positional encoding variant is also integrated. A trainable matrix $p_{learn} \in R^{T \times d_{proj}}$ is

initialized and updated through backpropagation, capturing spatial priors directly from the training data.

$$Z_t = H_t^{pos} + p_{learn}(t) \quad (51)$$

This dual embedding composed of deterministic sinusoidal encoding and adaptive learnable vectors, enables both global interpretability and fine-grained spatial alignment. The learnable vectors introduce flexibility in learning task-specific spatial sensitivities, such as edge proximity, region centrality, or band-specific transitions, which are especially relevant for adversarially perturbed or context-rich satellite images.

For multi-scale positional integration, an additional refinement block is applied across groups of tokens arranged by their patch origin in horizontal and vertical grids. For each token originating from a grid position (x, y), an auxiliary spatial encoding matrix $p_{grid} \in R^{(H/P)\times(W/P)d_{proj}}$ is defined. The corresponding vector $G_{x,y}$ is added as

$$Z'_{x,y} = Z_{x,y} + G_{x,y} \quad (52)$$

This grid-based augmentation embeds geometric structure over the patch grid. It strengthens the transformer's ability to learn region-specific behaviours, such as boundary transitions or terrain zoning, relevant for remote sensing classification. The inclusion of explicit 2D grid encodings further enhances the interpretability of attention heads when distributed across geographic subregions.

To stabilise the magnitude of position-encoded tokens and prevent dominance of any single component, normalization is performed using layer normalization

$$\hat{Z}_t = LayerNorm(Z'_t) \quad (53)$$

The normalized vector $\hat{Z}_t \in R^{d_{proj}}$ ensures uniform token scale before transformer attention begins, allowing each token to contribute equitably in the attention-weighted fusion process.

All normalized embeddings are collected into a sequence $\hat{Z} \in R^{T\times d_{proj}}$, which is passed directly into the transformer encoder. This sequence reflects the final token layout that encodes spectral information, spatial-locality, and adversarial resilience, unified through positional augmentation. The positional information encoding phase in CSAF-ViT bridges the abstraction gap between vision transformers and structured remote sensing imagery. Through a combination of sinusoidal, learnable, and grid-based encodings, the model learns to associate semantic importance with positional identity. This attention-guided spatial sensitivity forms the core structure for consistent, interpretable, and robust classification under varying adversarial and environmental conditions. Each token, now position-aware, enables the model to reason with awareness of where each semantic unit lies, ensuring alignment between spectral response, spatial behaviour, and global classification outcomes.

3.10. Transformer-Based Spatial Encoding

The position-augmented token sequence $\hat{Z} \in R^{T\times d_{proj}}$, which embeds both local spectral features and structured spatial identity, serves as the core input for transformer-based spatial encoding. The transformer module processes this token stream through a stack of attention layers, capturing global contextual dependencies between patch representations, irrespective of their physical adjacency. This mechanism enables the model to infer long-range spatial relationships critical for identifying extended structures such as roads, rivers, or contiguous vegetation patches under spectral noise or adversarial distortion.

Each transformer encoder layer consists of two primary sub-blocks: multi-head self-attention and position-wise feedforward transformation. The multi-head self-attention mechanism first maps the input tokens $\hat{Z}$ into query, key, and value representations using learnable projections $W^Q, W^K, W^V \in R^{d_{proj}\times d_h}$, where d_h denotes the dimensionality of each attention head. For a given token $\hat{Z}_t$, the projections are expressed as

$$Q_t = \hat{Z}_t \cdot W^Q, \quad K_t = \hat{Z}_t \cdot W^K, \quad V_t = \hat{Z}_t \cdot W^V \quad (54)$$

For a sequence of T tokens, the queries $Q \in R^{T\times d_h}$, keys $K \in R^{T\times d_h}$, and values $V \in R^{T\times d_h}$ are stacked across all tokens. Attention scores are computed as scaled dot products between each query and key, followed by softmax normalization

$$A_{ij} = \frac{Q_i \cdot K_j^T}{\sqrt{d_h}}, \quad \alpha_{ij} = \frac{exp(A_{ij})}{\sum_{k=1}^{T} exp(A_{ik})} \quad (55)$$

These attention weights α_{ij} define how much token i attends to the token j, establishing a dense interaction graph over the spatial layout of the image. The contextualised output for each token

is derived by aggregating value vectors based on these learned attention weights:

$$O_i = \sum_{j=1}^{T} \alpha_{ij} \cdot V_j \quad (56)$$

Multiple attention heads are instantiated in parallel to learn diverse aspects of spatial relations. Let h denote the number of heads; each head computes its contextual output $O_i^{(k)} \in R^{d_h}$, for $k = 1, \ldots, h$. These outputs are concatenated and projected back to the original dimension using a linear mapping $W^O \in R^{hd_h \times d_{proj}}$:

$$O_i^{concat} = Concat\left(O_i^{(1)}, \ldots, O_i^{(h)}\right), \; O_i^{final} = O_i^{concat} \cdot W^O \quad (57)$$

To stabilize the gradient flow and ensure representation coherence, a residual connection is applied by summing the original token embedding $\hat{Z}_i$ with the attention output:

$$R_i = \hat{Z}_i + O_i^{final} \quad (58)$$

Layer normalization is then used to prevent internal covariate shifts during optimization:

$$R_i^{norm} = LayerNorm(R_i) \quad (59)$$

The normalized outputs are passed through a position-wise feedforward network composed of two dense layers with an activation function $\delta(\cdot)$, typically GELU. This network enables nonlinear transformation of each token, refining spatial patterns and enhancing abstraction

$$F_i = W_2 \cdot \delta(W_1 \cdot R_i^{norm} + b_1) + b_2 \quad (60)$$

Here, $W_1 \in R^{d_{proj} \times d_{ff}}, W_2 \in R^{d_{ff} \times d_{proj}}$, and d_{ff} is the hidden layer size of the feedforward network. A second residual connection is added post-FFN, followed again by normalization:

$$T_i = LayerNorm(R_i^{norm} + F_i) \quad (61)$$

This process is repeated across all T tokens and stacked across L transformer layers. Each layer captures increasingly abstract relationships among spatial patches, gradually refining the representations through repeated global reasoning. The output after the final transformer layer is denoted by $T_{final} \in R^{T \times d_{proj}}$, where each token now encodes multi-level spatial correlations enriched through spectral modulation.

The classification token, placed at index $t = 0$, aggregates attention from all patches across all heads and layers. This token carries a comprehensive view of the global structure, with adversarial noise suppressed through spectrum-aware learning and global spatial encoding. The learned representation is later passed to a classification head for decision output. The transformer-based spatial encoding ensures that each patch embedding evolves into a globally contextualised vector. Attention heads focus not just on spatial proximity but also on long-range correlations informed by spectral alignment and structural variation.

3.11. CLS Token-Based Classification

The sequence of transformer-encoded tokens $T_{final} \in R^{T \times d_{proj}}$, where $T = N + 1$, contains one dedicated classification token and N patch tokens. Among them, the first token $T_0 \in R^{d_{proj}}$ is the classification-specific embedding, injected during the initial embedding phase and continuously refined across transformer layers. This special token aggregates attention-based signals from all spatial and spectral regions, making it a centralized semantic collector suitable for final category prediction.

To generate interpretable logits from the classification token, a linear projection is employed to map its representation into the output space. Let $W_{cls} \in R^{d_{proj} \times C}$ denote the learnable classification weight matrix, where C is the number of remote sensing classes. The unnormalized class scores $y \in R^C$ are computed by

$$y = T_0 \cdot W_{cls} + b_{cls} \quad (62)$$

The bias term $b_{cls} \in R^C$ allows for class-specific threshold shifts. The output vector y captures logit-level activations corresponding to each class, derived entirely from the spatially attentive and spectrally reweighted CLS token. This projection consolidates cross-patch evidence into a compact decision space.

To convert the logits into interpretable probabilities, a softmax activation is applied across the vector y, yielding normalized class probabilities $\hat{y} \in R^C$

$$\hat{y}_i = \frac{exp(y_i)}{\sum_{j=1}^{C} exp(y_j)} \quad (63)$$

Each element $\hat{y}_i$ represents the confidence assigned to class i. This output serves as the final prediction for downstream tasks, including land cover categorization, infrastructure detection, or anomaly segmentation in defence-oriented remote sensing workflows.

To enforce discriminative learning, a categorical cross-entropy loss is computed between the predicted probabilities and the ground truth label. $y^{true} \in R^C$, encoded as a one-hot vector. The classification loss L_{cls} is given as shown below

$$L_{cls} = -\sum_{i=1}^{C} y_i^{true} \cdot log(\hat{y}_i) \quad (64)$$

This loss encourages the model to maximize the likelihood of the correct class while penalizing misclassifications. The optimization process iteratively adjusts parameters in the transformer, spectral fusion layers, and attention blocks to minimize this value over the training dataset.

To reinforce robustness under adversarial perturbations and minor spectral shifts, an entropy-aware margin regularization is introduced. Let $y_{adv} \in R^C$ be the prediction on an adversarially perturbed version of the input. The divergence between clean and adversarial predictions is penalized through Kullback-Leibler divergence D_{KL}

$$L_{adv} = \sum_{i=1}^{C} \hat{y}_i \cdot log\left(\frac{\hat{y}_i}{y_{adv,i} + \epsilon}\right) \quad (65)$$

The small constant ϵ ensures numerical stability. The total objective function balances both clean classification performance and adversarial consistency.

$$L_{total} = L_{cls} + \gamma \cdot L_{adv} \quad (66)$$

Here, γ is a tunable scalar controlling the influence of adversarial consistency. The training loop minimizes L_{total} across batches, reinforcing the semantic reliability of the CLS token under both benign and adversarial environments.

To monitor feature separability and guide class boundary formation in latent space, a centre-based compactness regularizer is imposed. Let $\mu_i \in R^{d_{proj}}$ be the centroid of class i in the embedding space. For a sample belonging to the class i, the compactness loss L_{cmp} is formulated as

$$L_{cmp} = \|T_0 - \mu_i\|_2^2 \quad (67)$$

This metric penalizes deviation of the CLS token from the ideal class representation. Over time, the tokens representing the same class converge towards a common centre, leading to improved generalization and resilience against ambiguous or adversarial samples.

The CLS token serves not only as a classification unit but also as a carrier of interpretability. Attention-based inspection of the token's head-wise focus reveals its dependence on specific spatial patches and spectral regions. These attention maps can be visualized to trace decision patterns, enabling transparency in high-stakes applications such as defence surveillance or disaster response.

Throughout the classification phase, the network remains fully differentiable, allowing end-to-end training using standard backpropagation. The gradient flow spans from the classification token back through the transformer stack, patch embeddings, fusion modules, and spectral attention units, enabling coherent updates across the architecture.The transformer's ability to process global information ensures that the CLS token benefits from wide-ranging dependencies, unlike local receptive fields found in convolutional structures. This international field, enhanced by spectral alignment and robust encoding, positions the CLS token as the most information-dense and semantically aligned representation of the input scene. The final activation values derived from this token encapsulate compressed yet highly expressive summaries of the entire remote sensing instance, integrating multi-band textures, spatial boundaries, and contextual relationships into a unified prediction signal. The use of dedicated supervision mechanisms ensures that this representation is tuned not just for accuracy but also for interpretability and resilience in operational deployment scenarios.

3.12. Adversarial Training Integration

The classification predictions derived from the CLS token in CSAF-ViT are highly sensitive to input fidelity, particularly under perturbations crafted to exploit gradient-based vulnerabilities. To counter these threats, adversarial training is integrated into the learning loop, enabling the model to generalize under malicious distortions while preserving spectral-spatial integrity. This integration ensures that the attention-driven transformer does not collapse under spectral noise or subtle pixel-level manipulations.

Let the original remote sensing image be denoted as $I \in R^{H\times W\times C}$, where C is the number of spectral bands. A perturbation tensor $\delta \in R^{H\times W\times C}$is constructed to simulate adversarial noise, bounded by an l_∞-norm constraint.

$$\|\delta\|_\infty \leq \epsilon \tag{68}$$

The perturbed image $I^{adv} = I + \delta$ is introduced into the training loop, forcing the model to operate under challenging signal deviations. The adversarial tensor δ is iteratively optimized through Projected Gradient Descent (PGD) using gradients of the classification loss with respect to the input

$$\delta^{(t+1)} = \Pi_\epsilon\left[\delta^{(t)} + \alpha \cdot sign\left(\nabla_I L_{cls}\left(I^{(t)}, y^{true}\right)\right)\right] \tag{69}$$

Here, α is the step size, Π_ϵ denotes the projection operator that ensures the perturbation stays within the allowed norm ball, and $I^{(t)} = I + \delta^{(t)}$. This iterative scheme enables the generation of strong adversarial variants that retain visual similarity to clean inputs while inducing classification ambiguity.

To guide CSAF-ViT in resisting these perturbations, adversarial examples are passed through the entire pipeline, including spectral decomposition, attention-based encoding, transformer reasoning, and classification. The predicted probability vector $\hat{y}^{adv} \in R^C$ is compared against the ground truth label using the same categorical loss.

$$L_{adv}^{cls} = -\sum_{i-1}^{C} y_i^{true} \cdot log\left(\hat{y}_i^{adv}\right) \tag{70}$$

This adversarial loss penalizes performance degradation under hostile input shifts, pushing the model toward margin separation and entropy stabilisation. To balance learning between clean and adversarial inputs, a weighted joint loss is used, which is expressed as

$$L_{joint} = \lambda \cdot L_{cls} + (1-\lambda) \cdot L_{adv}^{cls} \tag{71}$$

The scalar $\lambda \in [0,1]$ controls the trade-off between natural classification and adversarial resistance. A curriculum schedule adjusts λ during training, initially favoring clean accuracy and gradually incorporating robustness as learning progresses.

To align internal representations under clean and adversarial variants, an embedding-level regularization is introduced. Let T_0 and T_0^{adv} denote the CLS tokens generated from clean and adversarial paths, respectively. A contrastive distance metric is applied.

$$L_{align} = \left\|T_0 - T_0^{adv}\right\|_2^2 \tag{72}$$

This alignment encourages the model to retain consistent semantic embeddings, regardless of input distortions. By regulating token-level drift, this loss ensures that attention distribution and spatial abstraction remain functionally equivalent across perturbed inputs.

To discourage overfitting to adversarial artefacts and ensure discriminative focus, a gradient-based attention sharpening loss is incorporated. Let $g_{i,j} \in R$ represent the gradient magnitude of the adversarial loss concerning the input pixel at position (i,j). The spectral-saliency sharpening penalty is defined as

$$L_{sharp} = \sum_{i,j}\left(g_{i,j} - \mu\right)^2 \tag{73}$$

Here, μ is the mean gradient magnitude over all spatial positions. This regularization pushes the gradient map toward a focused distribution, ensuring that only meaningful regions influence decision updates, rather than broad-spectrum noise.

The full adversarial training objective integrates classification, alignment, and attention sharpening

$$L_{total} = L_{joint} + \beta_1 \cdot L_{align} + \beta_2 \cdot L_{sharp} \tag{74}$$

Parameters $\beta_1$and β_2 determine the contribution of structural consistency and gradient focusing to the final objective. This multi-objective training loop strengthens CSAF-ViT's resilience by targeting both prediction stability and interpretive clarity.

Each gradient flow through this adversarial route is fully backpropagated across spectral reweighting units, attention heads, transformer blocks, and projection layers, ensuring that all modules learn to accommodate and neutralize noise. Spectral entropy modulation further interacts with these losses, amplifying attention toward robust high-information bands while suppressing low-informative or attack-prone regions. In practice, the adversarial integration is periodically alternated with natural training batches to avoid over-saturation. This alternating schedule ensures that the model learns to generalize on both clean and adversarial samples while avoiding mode

collapse. This dual-stream learning process shapes the parameters to form smooth, flat minima in the loss landscape, commonly associated with high robustness under distributional shifts.

The integrated adversarial strategy operationalized here enables CSAF-ViT to operate as a dependable remote sensing classification backbone, capable of resisting input tampering, spectral interference, and gradient-based attacks. Each module contributes to robustness not by isolation, but through a layered and tightly coupled response to adversarial variability embedded directly within the learning trajectory. Here is the PowerPoint slide content for CSAF-ViT across 12 slides, each containing exactly five crisp points, including one equation per slide with variable explanations and its role in the step. No dataset or result info is included.

4. DATASET DESCRIPTION

The EuroSat Dataset provides a rich archive of 27,000 RGB satellite images derived from Sentinel-2 multispectral observations. Each image is 64x64 pixels in size and represents one of 10 well-defined land use and land cover categories, including Annual Crop, Forest, Industrial, River, Residential, and more. Designed with uniform spatial resolution, the dataset ensures consistent feature extraction for machine learning models across diverse geographical regions in Europe. Captured under various seasonal and atmospheric conditions, the samples reflect natural variations in terrain and vegetation, adding robustness to classification tasks. The dataset is particularly suitable for benchmarking algorithms in land use detection, adversarial attack prevention, and geospatial scene classification. Each class is evenly distributed, mitigating class imbalance issues. By focusing on both urban and rural environments, the dataset captures essential patterns relevant to environmental monitoring and agricultural analysis. The diverse class composition and high inter-class variability offer a challenging ground for evaluating deep learning models' sensitivity to perturbations. Overall, EuroSat serves as a standard benchmark for training, validation, and testing of robust remote sensing models across cross-domain and adversarial settings.

5. RESULTS AND DISCUSSIONS

Results and discussions present a detailed comparative analysis of classification performance across different adversarial defense-enabled models tailored for remote sensing image classification. The primary evaluation emphasizes two critical performance indicators classification accuracy and F-measure both of which jointly reflect the reliability, precision, and robustness of feature representations and decision boundaries under adversarial interference. The comparison underscores how spectral integrity preservation, spatial context modeling, and attention-driven fusion influence classification stability.

Classification accuracy (CL-AC) denotes the proportion of correctly predicted labels, encompassing both true positives and true negatives, relative to the total number of test samples. This metric directly reflects the generalization strength of a model against both normal and perturbed samples. MSRF records a classification accuracy of 55.062%, pointing to moderate resistance against pixel-level distortions. A3OD improves this to 58.696%, owing to its integration of adaptive optimization strategies that support deeper spatial abstraction. CSAF-ViT achieves a notable accuracy of 62.613%, highlighting the effectiveness of its cross-spectral attention fusion and transformer-guided spatial encoding.

F-measure (F-MSR) evaluates the harmonic balance between precision and recall, offering insight into the model's capacity to maintain high detection fidelity while minimizing both false positives and false negatives. MSRF exhibits an F-measure of 55.952, indicating limited precision-recall balance under adversarial distortion. A3OD raises this to 59.721 by leveraging feature adaptivity and structural consistency. CSAF-ViT further outperforms with an F-measure of 63.949, showcasing strong semantic preservation and effective suppression of misleading gradients. The elevated F-measure reveals that CSAF-ViT minimizes false responses without compromising detection sensitivity, an essential requirement for remote sensing applications where misclassifications carry significant operational consequences.

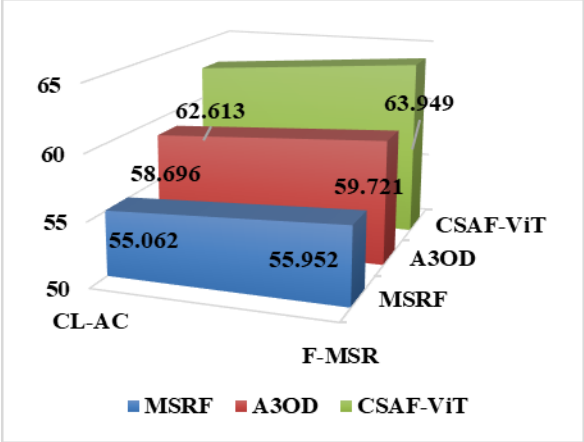

Figure 1. CL-AC and F-MSR

Fig 1 illustrates the observed performance trajectory in terms of both classification accuracy and F-measure, substantiating the superiority of CSAF-ViT in adversarially robust remote sensing classification.

Fowlkes Mallows Index (FMI) and Matthews Correlation Coefficient (MCC), two advanced metrics offering insights into classification reliability and correlation consistency, especially under adversarial disturbances. These metrics are essential in scenarios where class imbalance, feature distortions, or attack-based perturbations could otherwise mask model weaknesses not reflected by traditional accuracy scores. Fowlkes-Mallows Index assesses the geometric mean of precision and recall, quantifying the overlap between predicted and actual class clusters. A higher FMI indicates that the classifier maintains strong clustering behavior with minimal divergence between expected and predicted groupings. MSRF reaches an FMI score of 56.239, showing a baseline alignment with true class distributions. A3OD improves the FMI to 59.821, demonstrating better class cohesion through feature recalibration under adversarial influence. CSAF-ViT achieves a peak FMI of 64.123, indicating its ability to preserve both inter-class separation and intra-class compactness. This result confirms that the spectral-guided attention fusion embedded in CSAF-ViT not only refines local feature dependencies but also strengthens global structural preservation, minimizing misclassification noise.

Matthews Correlation Coefficient evaluates the quality of binary classifications by capturing the correlation between observed and predicted classifications. Unlike accuracy, MCC accounts for all four confusion matrix outcomes, TP, TN, FP, and FN offering a balanced view even under skewed class distributions. MSRF produces an MCC of 11.266, suggesting a weak correlation between predicted and ground-truth labels. A3OD improves this to 17.752 through optimized learning dynamics and spatial filtering. CSAF-ViT records the highest MCC of 25.859, confirming strong linear agreement and robust decision boundaries. The elevation in MCC reveals that the classifier effectively learns meaningful features that remain consistent even when adversarial noise is introduced.

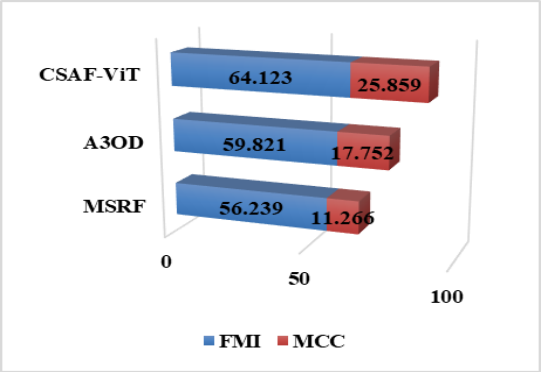

Figure 2. FMI and MCC

Fig 2. Illustrates the outcome of FMI and MCC in a pictorial format. Collectively, these trends further reinforce CSAF-ViT's dominant role in enabling robust classification with enhanced structural fidelity under attack-aware remote sensing scenarios.

Precision (PREC) and Recall (RCLL), which together evaluate the model's predictive confidence and sensitivity to true class identification under adversarial pressure. These metrics become vital in adversarial remote sensing scenarios where both false positives and false negatives can disrupt critical decisions in sensitive applications like disaster mapping or surveillance. Precision defines the proportion of true positives among all predicted positives, capturing how confidently the model assigns positive labels without overestimating their occurrence. MSRF yields a precision of 50.825%, reflecting lower reliability in distinguishing true class members from false predictions. A3OD improves the score to 56.460% by incorporating optimization-aware feature abstraction that filters misleading activations. CSAF-ViT leads with a precision of 59.562%, validating its strength in preserving spectral purity and spatial attention integrity across bands. The increase in precision indicates that

CSAF-ViT effectively suppresses irrelevant or adversarial cues that could mislead the classifier.

Recall measures the proportion of actual positives correctly identified, serving as a proxy for the model's completeness in detection. MSRF reaches a recall of 62.228%, while A3OD improves slightly to 63.383%, leveraging better region-level abstraction. CSAF-ViT records the highest recall of 69.032%, confirming its superior capacity to recover class-consistent features despite adversarial perturbations. This strong recall score reflects CSAF-ViT's ability to minimize missed detections by integrating multi-band semantics through transformer-guided reasoning.

Fig.3 exhibits the outcome of the models under the metrics PREC and RCLL. The performance gap observed across both precision and recall highlights CSAF-ViT's ability to maintain a balanced trade-off between false alarms and missed identifications. This balance is critical for remote sensing tasks, where overlooking key regions or falsely identifying patterns can significantly affect mission outcomes. These metrics solidify CSAF-ViT's role in producing resilient and context-aware classifications with reduced susceptibility to adversarial misguidance.

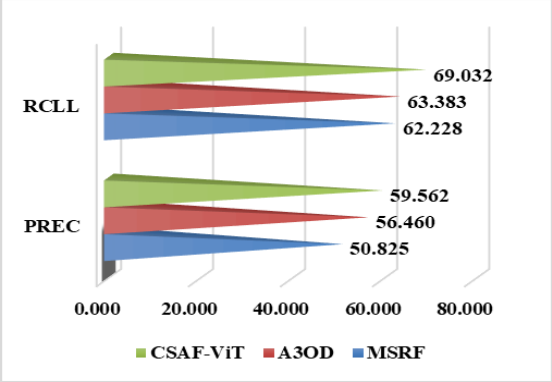

Figure 3. PREC and RCLL

6. CONCLUSION

The proposed CSAF-ViT model delivers a robust defense mechanism against adversarial perturbations in remote sensing image classification by integrating cross-spectral attention fusion with transformer-based spatial encoding. Through a twelve-stage architectural flow, the model systematically enhances spectral discrimination, spatial continuity, and structural integrity by addressing both inter-band inconsistencies and intraclass distortions. The model initiates with fine-grained spectral decomposition and progresses through band-wise feature encoding, attention-driven fusion, and transformer-guided classification all reinforced with adversarial training. Each component contributes to maximizing the retention of semantic fidelity under manipulated inputs. Empirical results demonstrate that CSAF-ViT surpasses baseline methods such as MSRF and A3OD across multiple evaluation metrics. Classification accuracy rises to 62.613%, with a corresponding F-measure of 63.949%, indicating improved balance between detection precision and recall. Furthermore, the model achieves the highest FMI of 64.123 and MCC of 25.859, validating its effectiveness in preserving meaningful feature correlations despite noise. Superior precision and recall values (59.562% and 69.032% respectively) underscore the model's capacity to maintain decision reliability across challenging spectral landscapes. Cross-spectral attention contributes significantly to preserving band-level coherence, while transformer encoding ensures contextual propagation of critical visual patterns across the fused representation. The combination effectively neutralizes gradient-based perturbation strategies by minimizing adversarial influence in both feature and decision spaces. Overall, CSAF-ViT presents a structurally resilient and interpretable architecture tailored for remote sensing security applications. Its integration of spectral-spatial alignment with adversarial robustness establishes a new direction for dependable classification frameworks in defense-oriented geospatial intelligence and environmental monitoring systems.

REFERENCES:

[1]. H. Feng et al., "Security of Target Recognition for UAV Forestry Remote Sensing Based on Multi-Source Data Fusion Transformer Framework", *Information Fusion*, vol. 112, 2024, p. 102555.

[2]. S. Sciancalepore, F. Davidovic, and G. Oligeri, "ORION: Verification of Drone Trajectories via Remote Identification Messages", *Future Generation Computer Systems*, vol. 160, 2024, pp. 869–878.

[3]. C. Yang, S. Wang, Y. Huang, and M. Guo, "SNR: One Single Network for Image Steganography with Robust Post-Save Recovery", *Neurocomputing*, vol. 651, 2025, p. 130929.

[4]. G. Zhou, L. Huang, and Q. Sun, "Fine-Grained Classification of Remote Sensing Ship Images Based on Improved VAN",

Computers, Materials and Continua, vol. 77, no. 2, 2023, pp. 1985–2007.

[5]. J. Nyengere *et al.*, "Forest Cover Restoration Analysis Using Remote Sensing and Machine Learning in Central Malawi", *Trees, Forests and People*, vol. 20, 2025, p. 100873.

[6]. Z. Ashkanani, R. Mohtar, M. Al-Momin, S. Hetrick, S. Al-Enezi, and N. Aladwani, "Integrating Neural Network Approaches with Remote Sensing for Detection and Prediction of Oil Contamination", *J Hazard Mater*, 2025, p. 139245.

[7]. T. Hussain, M. N. Khan, B. Yang, R. W. Attar, and A. Alhomoud, "LiDAR Point Cloud Transmission: Adversarial Perspectives of Spoofing Attacks in Autonomous Driving", *Comput Secur*, vol. 157, 2025, p. 104544.

[8]. P. Hemashree and G. Padmavathi, "Enhancing FGSM Attacks with Genetic Algorithms for Robust Adversarial Examples in Remote Sensing Image Classification Systems", in *Applications and Techniques in Information Security*, V. S. Shankar Sriram, A. G. H., G. Li, and S. R. Pokhrel, Eds., Singapore: Springer Nature Singapore, 2025, pp. 229–243.

[9]. Y. Xu, G. Xu, Z. An, M. H. Nielsen, and M. Shen, "Adversarial Attacks and Active Defense on Deep Learning Based Identification of GaN Power Amplifiers under Physical Perturbation", *AEU - International Journal of Electronics and Communications*, vol. 159, 2023, p. 154478.

[10]. S. B. U. Haque, "A Fuzzy-Based Frame Transformation to Mitigate the Impact of Adversarial Attacks in Deep Learning-Based Real-time Video Surveillance Systems", *Appl Soft Comput*, vol. 167, 2024, p. 112440.

[11]. Y. Lu, "LSEVGG: An Attention Mechanism and Lightweight-Improved VGG Network for Remote Sensing Landscape Image Classification", *Alexandria Engineering Journal*, vol. 127, 2025, pp. 943–951.

[12]. H. Yang *et al.*, "A High-Resolution Remote Sensing Land Use/Land Cover Classification Method Based on Multi-Level Features Adaptation of Segment Anything Model", *International Journal of Applied Earth Observation and Geoinformation*, vol. 141, 2025, p. 104659.

[13]. Y. Xu *et al.*, "Construction and Application of a Drought Classification Model for Tea Plantations Based on Multi-Source Remote Sensing", *Smart Agricultural Technology*, vol. 12, 2025, p. 101132.

[14]. C. Li *et al.*, "MIFFNet: Feature Fusion-Oriented Classification of Volcanic Lithology from Remote Sensing Image", *Alexandria Engineering Journal*, vol. 115, 2025, pp. 538–552.

[15]. X. Li, J. Li, J. Jiang, X. Pan, and X. Huang, "Spatio-Temporal-Text Fusion for Hierarchical Multi-Label Crop Classification Based on Time-Series Remote Sensing Imagery", *International Journal of Applied Earth Observation and Geoinformation*, vol. 139, 2025, p. 104471.

[16]. G. Zhou, L. Huang, and Q. Sun, "Fine-Grained Classification of Remote Sensing Ship Images Based on Improved VAN", *Computers, Materials and Continua*, vol. 77, no. 2, 2023, pp. 1985–2007.

[17]. M. Jaber, S. Elmi, M. Nassar, and W. El Hajj, "Introducing Residual Networks to Vision Transformers for Adversarial Attacks", *Procedia Comput Sci*, vol. 246, 2024, pp. 423–432.

[18]. H. Yin, Y. Liu, Y. Li, Z. Guo, and Y. Wang, "Defeating Deep Learning Based De-Anonymization Attacks with Adversarial Example", *Journal of Network and Computer Applications*, vol. 220, 2023, p. 103733.

[19]. B. Peng, B. Peng, J. Xia, T. Liu, Y. Liu, and L. Liu, "Towards Assessing the Synthetic-to-Measured Adversarial vulnerability of SAR ATR", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 214, 2024, pp. 119–134.

[20]. Ö. Sen *et al.*, "Simulation of Multi-Stage Attack and Defense Mechanisms in Smart Grids", *International Journal of Critical Infrastructure Protection*, vol. 48, 2025, p. 100727.

[21]. P. Zhu *et al.*, "Transferable Adversarial Attacks for Multi-Model Systems Coupling Image Fusion with Classification Models", *Cybersecurity*, vol. 8, no. 1, 2025, p. 23.

[22]. L. Tang, Z. Yin, H. Su, W. Lyu, and B. Luo, "WFSS: Weighted Fusion of Spectral Transformer and Spatial Self-Attention for

Robust Hyperspectral Image Classification Against Adversarial Attacks", *Visual Intelligence*, vol. 2, no. 1, 2024, p. 5.

[23]. A. D. M. Ibrahum, M. Hussain, and J.-E. Hong, "Deep Learning Adversarial Attacks and Defenses in Autonomous Vehicles: a Systematic Literature Review from a Safety Perspective", *Artif Intell Rev*, vol. 58, no. 1, 2024, p. 28.

[24]. R. Teixeira, M. Antunes, J. P. Barraca, D. Gomes, and R. L. Aguiar, "Rethinking Security: the Resilience of Shallow ML Models", *Int J Data Sci Anal*, 2024.

[25]. J. Liu, Y. Feng, X. Liu, J. Zhao, and Q. Liu, "MRm-DLDet: a Memory-Resident Malware Detection Framework Based on Memory Forensics and Deep Neural Network," *Cybersecurity*, vol. 6, no. 1, 2023, p. 21.

[26]. H. Fan and G. Wei, "Multi-Spectral Remote Sensing Image Fusion Method Based on Gradient Moment Matching", *Systems and Soft Computing*, vol. 6, 2024, p. 200108.

[27]. R. Huang, L. Chen, J. Zheng, Q. Zhang, and X. Yu, "Adversarial Attacks Against Object Detection in Remote Sensing Images", in *Artificial Intelligence Security and Privacy*, J. Vaidya, M. Gabbouj, and J. Li, Eds., Singapore: Springer Nature Singapore, 2024, pp. 358–367.

[28]. J. Ramkumar, K. S. Jeen Marseline, and D. R. Medhunhashini, "Relentless Firefly Optimization-Based Routing Protocol (RFORP) for Securing Fintech Data in IoT-Based Ad-Hoc Networks," Int. J. Comput. Networks Appl., vol. 10, no. 4, pp. 668–687, 2023, doi: 10.22247/ijcna/2023/223319.

[29]. M. P. Swapna and J. Ramkumar, "Multiple Memory Image Instances Stratagem to Detect Fileless Malware," in Communications in Computer and Information Science, S. Rajagopal, K. Popat, D. Meva, and S. Bajeja, Eds., Cham: Springer Nature Switzerland, 2024, pp. 131–140. doi: 10.1007/978-3-031-59100-6_11.

[30]. R. Jaganathan, S. Mehta, and R. Krishan, Bio-Inspired intelligence for smart decision-making, vol. i. 2024. doi: 10.4018/9798369352762.

[31]. S. P. Geetha, N. M. S. Sundari, J. Ramkumar, and R. Karthikeyan, "Energy Efficient Routing In Quantum Flying Ad Hoc Network (Q-FANET) Using Mamdani Fuzzy Inference Enhanced Dijkstra's Algorithm (MFI-EDA)," J. Theor. Appl. Inf. Technol., vol. 102, no. 9, pp. 3708–3724, 2024, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85197297302&partnerID=40&md5=72d51668bee6239f09a59d2694df67d6

[32]. R. Vadivel and J. Ramkumar, "QoS-enabled improved cuckoo search-inspired protocol (ICSIP) for IoT-based healthcare applications," in Incorporating the Internet of Things in Healthcare Applications and Wearable Devices, IGI Global, 2019, pp. 109–121. doi: 10.4018/978-1-7998-1090-2.ch006.

[33]. J. Ramkumar and R. Vadivel, "Improved Wolf prey inspired protocol for routing in cognitive radio Ad Hoc networks," Int. J. Comput. Networks Appl., vol. 7, no. 5, pp. 126–136, 2020, doi: 10.22247/ijcna/2020/202977.

[34]. D. Jayaraj, J. Ramkumar, M. Lingaraj, and B. Sureshkumar, "AFSORP: Adaptive Fish Swarm Optimization-Based Routing Protocol for Mobility Enabled Wireless Sensor Network," Int. J. Comput. Networks Appl., vol. 10, no. 1, pp. 119–129, 2023, doi: 10.22247/ijcna/2023/218516.

[35]. R. Jaganathan, S. Mehta, and R. Krishan, "Preface," Bio-Inspired Intell. Smart Decis., pp. xix–xx, 2024, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195725049&partnerID=40&md5=7a2aa7adc005662eebc12ef82e3bd19f

[36]. R. Jaganathan, S. Mehta, and R. Krishan, "Preface," Intell. Decis. Mak. Through Bio-Inspired Optim., pp. xiii–xvi, 2024, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192858710&partnerID=40&md5=f8f1079e8772bd424d2cdd979e5f2710

[37]. J. Ramkumar, R. Karthikeyan, and K. O. Nitish, "Securing Library Data With Blockchain Advantage," in Enhancing Security and Regulations in Libraries with Blockchain Technology, 2024, pp. 117–138. doi: 10.4018/979-8-3693-9616-2.ch006.

[38]. P. S. Ponnukumar, N. I. Francis Xavier, and R. Jaganathan, "Stable Plithogenic Cubic Sets," J. Fuzzy Ext. Appl., vol. 6, no. 2, pp. 410–423, 2025, doi: 10.22105/jfea.2025.449408.1422.

[39]. V. Valarmathi and J. Ramkumar, "Modernizing Wildfire Management Through Deep Learning and IoT in Fire Ecology," in Machine Learning and Internet of Things in Fire Ecology, 2024, pp. 203–229. doi: 10.4018/979-8-3693-7565-5.ch0010.

[40]. B. Suchitra, R. Karthikeyan, J. Ramkumar, and V. Valarmathi, "Enhancing Recurrent Neural Network Performance for Latent Autoimmune Diabetes Detection (Lada) Using Exocoetidae Optimization," J. Theor. Appl. Inf. Technol., vol. 103, no. 5, pp. 1645–1667, 2025, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105000948603&partnerID=40&md5=66c8f111b153fed68b3d0ea9c88c411e

[41]. S. P. Priyadharshini, F. Nirmala Irudayam, and J. Ramkumar, "An Unique Overture of Plithogenic Cubic Overset, Underset and Offset," in Studies in Fuzziness and Soft Computing, vol. 435, 2025, pp. 139–156. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105001675443&doi=10.1007%2F978-3-031-78505-4_7&partnerID=40&md5=e9def8c6a233de4fbf8f1549ad72027f

[42]. M. Lingaraj, T. N. Sugumar, C. S. Felix, and J. Ramkumar, "Query aware routing protocol for mobility enabled wireless sensor network," Int. J. Comput. Networks Appl., vol. 8, no. 3, pp. 258–267, 2021, doi: 10.22247/ijcna/2021/209192.