

INRA-Transformer: A Multi-Scale Hybrid Vision Transformer for Precise Seismic Deformation Mapping of the 6.1 Mw Sındırgı Earthquake

İlyas ASLAN

ilyas.aslan@dicle.edu.tr

Dicle University

Research Article

Keywords: Differential Synthetic Aperture Radar (DInSAR), Vision Transformers (ViT), Hybrid deep learning, Seismic deformation mapping, Sındırgı earthquake, Inception module, Explainable AI (XAI)

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10777482/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Precise mapping of surface deformation is essential for seismic hazard assessment and tectonic process interpretation. Differential Interferometric Synthetic Aperture Radar (DInSAR) is an established geodetic technique for detecting ground displacement; however, conventional classifiers and standard deep learning models often struggle to resolve subtle deformation signals from complex phase noise and decorrelation artifacts. This study introduces INRA-Transformer, a novel hybrid Vision Transformer architecture designed to enhance post-seismic deformation mapping by integrating multi-scale feature extraction with global contextual reasoning. Focusing on the 6.1 Mw Sındırgı earthquake (Balıkesir, Türkiye), Sentinel-1A DInSAR interferograms were processed to evaluate model performance in a real-world seismic scenario. Unlike standard ViTs, the INRA-Transformer incorporates an Inception-based multi-scale patch embedding stem and a depthwise-convolutional feed-forward network (DW-FFN) to inject local inductive bias into the Transformer framework. Experimental results demonstrate that the proposed architecture achieves an overall accuracy (OA) of 97.58% and a Kappa coefficient of 0.966 across 50,854 validation samples. Class-wise evaluation highlights exceptional performance in identifying high-intensity deformation zones ($F1\text{-score} = 0.99$). Furthermore, Explainable AI (XAI) attention mapping confirms that the network effectively suppresses decorrelation noise and atmospheric phase screens by prioritizing physically consistent phase gradients. These findings demonstrate that embedding multi-scale convolutional priors within Vision Transformers provides a robust, interpretable mechanism for rapid geodetic assessment and earthquake monitoring.

1. Introduction

The precise mapping of surface deformations is a fundamental prerequisite for assessing seismic hazards and understanding the complex geodynamics of active fault systems. Differential Interferometric Synthetic Aperture Radar (DInSAR) has emerged as the premier geodetic tool for monitoring these displacements, offering millimetric precision over vast spatial swaths regardless of atmospheric illumination (Bakon et al., 2020) (Zhang et al., 2019). By calculating the phase difference between multi-temporal SAR acquisitions, DInSAR allows for the quantification of Line-of-Sight (LOS) surface changes, providing a critical data source for analyzing significant tectonic events such as the 6.1 Mw Sındırgı earthquake (Balıkesir, Türkiye).

To translate these complex interferometric phase signals into actionable geodetic maps, researchers have historically relied on supervised classification techniques (Aslan & Polat, 2024). In the initial stages of automated mapping, the Maximum Likelihood Classification (MLC) algorithm served as a standard approach for generating ground-truth data and delineating deformation zones based on statistical probability distributions (Wassie et al., 2025). While MLC provided a foundational framework for remote sensing classification, it remains constrained by its reliance on Gaussian distribution assumptions and its inability to account for the spatial context and non-linear noise patterns (e.g., atmospheric phase screens and decorrelation) inherent in Sentinel-1A imagery.

The limitations of statistical classifiers led to a paradigm shift toward Deep Learning (DL). Convolutional Neural Networks (CNNs) revolutionized the field by introducing hierarchical feature extraction, allowing models to automatically learn spatial textures and fringe patterns without manual feature engineering (Zhu et al., 2021). However, as highlighted in recent studies, CNNs are inherently "local" operators; their receptive fields are restricted by kernel sizes, which often prevents them from capturing the long-range spatial dependencies and broad tectonic trends that define large-scale seismic displacements (Dosovitskiy et al., 2021).

The next evolutionary step in this technological trajectory is the Vision Transformer (ViT). Unlike previous architectures, Transformers utilize self-attention mechanisms to establish global relationships between distant pixels, effectively treating the entire interferogram as a connected system. This is particularly vital for seismic analysis, where the deformation gradient at the epicenter is physically linked to stable regions kilometers away. Despite their global reasoning power, standard ViTs lack the "inductive bias" required to recognize the fine-grained, micro-scale phase transitions that define precise LOS displacement boundaries.

The automation of DInSAR analysis has undergone a significant paradigm shift with the advent of Deep Learning (DL). Initial efforts primarily focused on the use of CNNs for detecting and classifying deformation patterns. Kumar & Devanaboyina, (2026) demonstrated the efficacy of CNNs in identifying global volcanic deformation signals in Sentinel-1 data, proving that DL could outperform manual inspection in large datasets. Similarly, Zhou et al. (2021) provided a comprehensive review of DL-based phase unwrapping, noting that residual networks significantly reduce the errors caused by phase discontinuities compared to traditional mathematical solvers.

Despite the localized success of CNNs, their limited receptive field often restricts their ability to model the long-range spatial dependencies inherent in broad tectonic displacements. To address this, attention-based mechanisms have been introduced to the remote sensing domain. Zhu et al. (2021) emphasized that while CNNs are "inductive bias heavy," they lack the global contextual reasoning required for complex fringe interpretation. This led to the adoption of Vision Transformers (ViT). Sun et al. (2022) and Ni et al. (2024) showcased that Transformers could achieve superior performance in SAR image classification by leveraging self-attention to model relationships between distant pixels, effectively capturing the "big picture" of the deformation field.

Most recently, the literature has moved toward hybrid architectures to balance local and global feature extraction. He et al. (2024) proposed a hybrid CNN-Transformer network for SAR change detection, showing that convolutional "stems" act as an excellent feature-enhancer for Transformer encoders. Zhou et al. (2025) introduced a multi-scale Inception-based Transformer for seismic signal identification, highlighting that deformation signals exist at multiple spatial frequencies. In the context of phase gradient analysis, Ross et al. (2024) utilized deep learning on DInSAR data to automate grounding line delineation, demonstrating that DL-based phase gradient analysis can reach expert-level precision. Further contributions by Wei et al. (2024) and Xie, (2025) have explored machine learning for landslide

and earthquake mapping, yet a gap remains in architectures that specifically treat wrapped phase gradients as multi-scale features.

In this context, our study introduces the INRA-Transformer, a hybrid architecture designed to bridge the gap between multi-scale convolutional priors and global attention mechanisms. By focusing on the Sındırgı earthquake, we contribute to the literature in three significant ways:

1. **Methodological Synergy:** We propose a novel "Inception-patch stem" that captures multi-scale features, combined with a "Depthwise-FFN" that maintains spatial integrity, overcoming the resolution limitations of standard ViTs.
2. **High-Precision Quantification:** We demonstrate that our model can classify complex deformations with an overall accuracy of 97.58% and a Kappa coefficient of 0.966 within a specific pm 0.03 m displacement scale.
3. **Explainability and Robustness:** Through XAI-based attention mapping, we provide a geophysically consistent validation, proving that the model suppresses decorrelation noise to focus on true tectonic gradients—a level of interpretability often missing in previous DL-based DInSAR studies.

2. Material and Methods

2.1. Study Area and Seismic Event

The study focuses on the 6.1 Mw Sındırgı earthquake located in the Balıkesir province of western Türkiye. This region is governed by an active extensional and strike-slip tectonic regime within the Western Anatolian Extensional Province (Shafapourtehrany et al., 2023). The seismic event induced significant Line-of-Sight (LOS) surface displacements across complex topographic terrain, presenting a challenging environment for evaluating model robustness against spatial decorrelation and atmospheric phase screens (Xie, 2025). The geographical context and the main fault systems of the study area are illustrated in Fig. 1.

2.2. Data Source and DInSAR Processing

Differential interferograms were generated from C-band Sentinel-1A Synthetic Aperture Radar (SAR) acquisitions captured in Interferometric Wide (IW) Single Look Complex (SLC) mode. The interferometric pair covers the pre-seismic and post-seismic periods of the Sındırgı event. Topographic phase components were removed using the 30-meter Shuttle Radar Topography Mission (SRTM) DEM, and adaptive phase filtering was applied to optimize coherence. The resulting wrapped phase interferogram, highlighting the seismic deformation fringes, is presented in Fig. 2.

The interferogram is the core input for our model, where the color fringes represent the surface deformation field (Albino & Biggs 2021). These fringes are subject to various noise sources, such as temporal decorrelation and atmospheric artifacts, which the INRA-Transformer is designed to filter out by focusing on geophysically consistent phase gradients (Ross et al., 2024).

2.3. Physical Scaling and Ground Truth Generation

To create a supervised learning environment, the interferogram was labeled into discrete classes using a combined approach of statistical estimation and expert validation.

Ground Truth Generation: Initially, the Maximum Likelihood Classification (MLC) algorithm was used to categorize the phase values (Li et al., 2023). These results were then manually refined based on the technical legends to ensure the accuracy of the training labels.

Displacement Scaling: The labeling process was governed by the physical parameters of the C-band SAR signal. As shown in the technical legends (Fig. 3), the dataset is scaled to reflect a Line-of-Sight (LOS) displacement range of approximately -0.03 m (subsidence) to $+0.03$ m (uplift). In the wrapped phase domain, each full fringe cycle ($-\pi$ to $+\pi$) corresponds to 2.8 cm of ground movement (Daramola et al., 2025).

3. Proposed INRA-Transformer Architecture

3.1. Vision Transformer Foundation

The Vision Transformer (ViT), originally introduced by Dosovitskiy et al. (2021), represents a significant departure from the pixel-to-pixel processing logic of Convolutional Neural Networks (CNNs). Unlike CNNs, which rely on localized kernels to extract features, ViTs treat an image as a sequence of patches, applying the Self-Attention mechanism to capture global relationships across the entire input (Zhu et al., 2021).

In the context of remote sensing and DInSAR, the self-attention mechanism provides a distinct advantage: it allows the model to compute the correlation between distant pixels in an interferogram. This is geophysically critical because tectonic movements at the earthquake epicenter are mathematically and physically linked to stable reference points situated kilometers away (Wang et al., 2023). However, standard ViTs are often "data-hungry" and lack the inductive bias (such as translation invariance and locality) that is inherent in convolutions. Recent studies suggest that while ViTs excel at global reasoning, they may struggle with the fine-grained, high-frequency phase gradients found in wrapped interferograms unless they are supplemented with local feature extractors (Sun et al., 2022; Zhou et al., 2025).

3.2. Inception Depthwise Separable Convolution Block

Inception Depthwise Separable Convolution is a concept that combines two powerful ideas in deep learning architectures: inception modules and depthwise separable convolutions (Fig. 4). The goal is to create a more efficient and lightweight architecture while preserving the multi-scale feature extraction ability of Inception modules. The inception module is designed to extract features at multiple scales by using different sizes of convolutional kernels (e.g., 1×1 , 3×3 , 5×5) and pooling layers in parallel. These

features are concatenated at the output, making the module capable of learning diverse spatial information. The depthwise separable convolution is a factorized convolution operation that splits the standard convolution into two steps: Depthwise convolution (DC) applies a single filter to each input channel independently. Pointwise convolution (PC) applies a 1x1 convolution to create a linear combination of the output channels. This factorization significantly reduces the number of parameters and computations compared to a standard convolution, making the model more efficient. The Inception depthwise separable convolution module uses depthwise separable convolutions in place of standard convolutions within the inception module. Similar to the inception module, parallel branches process the input using depthwise separable convolutions with varying kernel sizes (e.g., 3x3, 5x5). Each branch captures spatial features at different scales while reducing computational overhead. The outputs from all branches are concatenated along the channel dimension to form the final feature map.

3.3. Proposed INRA-Transformer Architecture

To address the limitations of standard ViTs, we propose the INRA-Transformer, a hybrid model that integrates convolutional layers into the Transformer framework. The architecture is composed of three main components designed to handle the noise and scale complexity of the Sındırgı earthquake dataset:

- **Inception-Based Patch Embedding (Multi-Scale Stem):**Standard ViTs use a single-scale linear projection to divide the image into patches. However, seismic signals exist at multiple spatial scales. Our model replaces the linear projection with an Inception-style stem (Zhou et al., 2025). By using parallel convolutional filters (e.g., 3x3, 5x5, 7x7), the model can capture micro-scale phase transitions while simultaneously identifying broad deformation trends. This multi-scale approach is particularly effective at suppressing the chaotic decorrelation noise often found in Sentinel-1A data (Daramola et al., 2025).
- **Transformer Encoder with Global Attention:** The core of the architecture consists of multiple Transformer blocks. Each block employs Multi-Head Self-Attention (MHSA) to model the global context of the displacement field. Following the principles of Ross et al. (2024), this global perspective ensures that the model differentiates between regional tectonic signals and localized atmospheric artifacts, which might otherwise be misclassified by purely local CNNs.
- **Depthwise-Convolutional Feed-Forward Network (DW-FFN):**Instead of the standard position-wise MLP found in traditional Transformers, we implement a Depthwise-Convolutional FFN (Wassie et al., 2025). This modification reintroduces a local spatial correlation into the sequence processing, ensuring that the spatial integrity and physical continuity of the interferometric fringes are preserved throughout the network layers.

The complete schematic of the proposed INRA-Transformer architecture is depicted in Fig. 5.

4. Experimental Results

4.1. Evaluation criteria

In the experimental evaluation, six quantitative metrics were utilized to assess the classification performance of the proposed INRA-Transformer architecture and compare it with benchmark models. These metrics include Overall Accuracy (OA), Average Accuracy (AA), Kappa coefficient (K), Precision, Recall, and F1-score. All evaluation criteria are derived directly from the pixel-level confusion matrix generated over the validation dataset. In the context of DInSAR deformation mapping, the confusion matrix elements (TP, TN, FP, FN) are defined as follows:

True Positives (TP): The number of pixels correctly predicted by the model as belonging to a specific deformation class (e.g., correctly classifying a pixel as High deformation).

True Negatives (TN): The number of pixels correctly identified as not belonging to a specific deformation class.

False Positives (FP): The number of pixels incorrectly assigned to a specific deformation class when they actually belong to another category (e.g., misclassifying a Stable pixel as Low deformation).

False Negatives (FN): The number of pixels belonging to a specific deformation class that the model failed to detect. The mathematical formulations for these evaluation metrics are defined as follows:

$Average Accuracy (AA) = \frac{\sum_{i=1}^N Accuracy(N_{ClassN})}{N} \quad (2)$
$Overall Accuracy (OA) = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$
$Kappa (K) = \frac{P_o - P_e}{1 - P_e} \quad (4)$
$P_o = \frac{Total\ number\ of\ correctly\ classified\ instances (TP+TN)}{Total\ number\ of\ instances (TP+TN+FP+FN)}$
$P_e = \sum_i \frac{n_{ii} \cdot n_{+i}}{N^2}$
$Precision = \frac{TP}{TP+FP} \quad (5)$
$Recall = \frac{TP}{TP+FN} \quad (6)$
$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$

In Eq. (3), P_o represents the observed proportion of agreement between the predicted deformation map and the ground truth labels. P_e denotes the expected proportion of agreement due to chance. N_{i+} is the total number of pixels predicted as class i (row sum in the confusion matrix), n_{+i} is the total number of

actual ground-truth pixels in class i (column sum in the confusion matrix), and N represents the total validation pixel count ($N = 50,854$). The F1-score serves as a critical balanced metric, particularly in DInSAR spatial analysis where class volumes vary between high-intensity rupture cores and broad stable background areas.

4.2. Classification Performance Analysis

The quantitative classification performance of the proposed INRA-Transformer architecture was thoroughly evaluated using the validation dataset derived from the Sentinel-1A DInSAR interferogram. The primary objective is to differentiate various seismic displacement intensities and stable background regions across the 6.1 Mw Sındırgı earthquake area. To comprehensively evaluate the pixel-level classification capability of our hybrid framework, a total of 50,854 validation pixels were analyzed across five distinct deformation categories.

The confusion matrix obtained from the experimental studies conducted with the proposed method provides a detailed breakdown of correct and misclassified pixel counts across all classes. The resulting class-wise distribution and prediction accuracy are visualized in the confusion matrix in Fig. 6.

Analyzing the class-wise pixel distribution from the confusion matrix reveals the following detailed classification performance:

Class 1 (Low deformation): Out of 9,664 actual Low deformation pixels, 9,202 samples were correctly classified into their true category. Among the misclassified samples, 318 pixels were misclassified as Class 3 (Stable), 133 pixels as Class 4 (Moderate deformation), and 11 pixels as Class 5 (High deformation).

Class 2 (Very high deformation): Out of 3,219 actual Very high deformation pixels, 3,095 samples were correctly identified. Of the remaining misclassified pixels, 80 were assigned to Class 5 (High deformation), 43 to Class 4 (Moderate deformation), and only a single pixel (1 sample) was misclassified as Class 1 (Low deformation).

Class 3 (Stable): Out of 15,136 actual non-deformed background (Stable) samples, 14,887 pixels were correctly categorized. For the erroneous predictions, 163 samples were misclassified as Class 1 (Low deformation), 80 samples as Class 5 (High deformation), and 6 samples as Class 4 (Moderate deformation).

Class 4 (Moderate deformation): Out of 2,290 actual Moderate deformation samples, 2,092 pixels were accurately classified. Among the misclassified instances, 121 pixels were attributed to Class 1 (Low deformation), 61 pixels to Class 5 (High deformation), 9 pixels to Class 3 (Stable), and 7 pixels to Class 2 (Very high deformation).

Class 5 (High deformation): Representing the primary seismic rupture zones with intense phase gradients, 20,349 samples were correctly classified out of 20,545 actual High deformation pixels. The

minor misclassifications included 72 pixels classified as Class 4 (Moderate deformation), 68 pixels as Class 3 (Stable), 43 pixels as Class 2 (Very high deformation), and 13 pixels as Class 1 (Low deformation).

The vast majority of the predictions fall directly along the primary diagonal of the confusion matrix, demonstrating the exceptional discriminative power of combining multi-scale Inception convolutional stems with global self-attention mechanisms. The minor misclassifications predominantly occur between spatially adjacent deformation boundaries, which is geophysically expected due to continuous phase transitions and localized atmospheric phase screen (APS) residual noise in DInSAR processing.

To comprehensively evaluate the performance of our hybrid INRA-Transformer architecture on the Sentinel-1A DInSAR validation dataset, class-based accuracy, precision, recall, F1-score, as well as general performance indicators including Overall Accuracy (OA), Average Accuracy (AA), and Kappa coefficient (K) were computed. The quantitative classification metrics derived from the validation experiments across all five deformation categories are presented in Table 1.

Table 1
Detailed classification results of the proposed INRA-Transformer model (%)

Evaluation Metrics	Class 1 (Low deformation)	Class 2 (Very high deformation)	Class 3 (Stable)	Class 4 (Moderate deformation)	Class 5 (High deformation)
Accuracy	95.22	96.15	98.36	91.35	99.05
Precision	96.86	98.41	97.42	89.17	98.87
Recall	95.22	96.15	98.36	91.35	99.05
F1-score	96.03	97.27	97.89	90.25	98.96
Overall Accuracy (OA)	97.58%				
Average Accuracy (AA)	96.03%				
Kappa Coefficient (K)	96.60%				

To analyze the performance of our hybrid method on the DInSAR validation dataset, class-based accuracy, precision, recall, F1-score, OA, AA, and K evaluation metrics were calculated (Table 1). Among the five displacement classes, our proposed INRA-Transformer method classified Class 5 (High deformation) with the highest accuracy of 99.05% and an F1-score of 98.96%. Furthermore, the Class 3 (Stable) background zone achieved a 98.36% accuracy and 97.42% precision, while Class 2 (Very high

deformation) reached 96.15% accuracy. The Class 1 (Low deformation) and Class 4 (Moderate deformation) classes yielded accuracies of 95.22% and 91.35%, respectively.

Looking at the overall evaluation criteria across all 50,854 validation pixels, the proposed model achieved an Overall Accuracy (OA) of 97.58%, an Average Accuracy (AA) of 96.03%, and a Kappa coefficient (K) of 96.60%. The lowest class performance was observed in Class 4 (Moderate deformation) due to subtle phase transition gradients and atmospheric noise screens. Nevertheless, these high overall metrics confirm that the integration of Inception-based multi-scale convolutional stems and Depthwise separable FFN layers within a Vision Transformer framework effectively addresses the complex spatial and scale challenges inherent in C-band Sentinel-1A interferograms.

In Fig. 7, you can observe the accuracy and loss convergence of our suggested INRA-Transformer approach when applied to training sets spanning 100 epochs. Upon scrutinizing Fig. 7, it becomes evident that training accuracy steadily rises throughout the 100 epochs, while the loss consistently diminishes. Notably, rapid convergence is attained after approximately 30 epochs, and the training curve stabilizes near epoch 70, underscoring the rapid convergence and training stability exhibited by our method without showing signs of overfitting.

The spatial prediction map generated by the proposed INRA-Transformer model is compared with the ground truth map in Fig. 8. The color-coded deformation scale ranges from the central core of maximum surface shift to the outer un-deformed background:

Very High Deformation (Blue): Highlights the epicenter's rupture core where the primary phase gradient is highest.

High Deformation (White): Captures the immediate inner ring around the epicenter with strong fringe continuity.

Moderate Deformation (Brown): Delineates the middle transition boundary of surface movement.

Low Deformation (Dark Green): Represents the outer edge of subtle displacement.

Stable (Yellow): Covers the non-deformed regional background.

As observed in Fig. 8, the proposed model displays an exceptionally high spatial agreement with the ground truth map, cleanly isolating the blue core of maximum deformation without producing blocky pixel noise.

To further interpret the internal decision-making process of the proposed INRA-Transformer model, Explainable AI (XAI) attention maps were extracted from the final Transformer encoder layer. Figure 9 illustrates the spatial distribution of attention weights across the Sındırgı earthquake interferogram.

As illustrated in Fig. 9, the attention maps reveal that the self-attention mechanism dynamically assigns the highest feature weights (represented by warm/intense region highlights) directly to the high-gradient

phase fringes located at the earthquake epicenter.

Physical Fringe Prioritization: Rather than spreading weights uniformly or getting distracted by background pixels, the network explicitly concentrates its focus on the primary rupture boundaries. This proves that the Inception-based multi-scale patch embedding stem successfully directs the Transformer layers toward geophysically significant phase transitions.

Noise Suppression: Regions affected by temporal decorrelation and atmospheric phase screens (APS) exhibit noticeably low attention weights. This indicates that the hybrid architecture acts as an implicit spatial filter, suppressing un-correlated noise while amplifying true tectonic signals. This explainability analysis confirms that the state-of-the-art accuracy achieved by the INRA-Transformer (97.58% OA) is driven by physically consistent feature learning rather than shortcut learning or overfitting.

4.3. Comparative Analysis: INRA-Transformer vs. Standard Vision Transformer (ViT)

To rigorously validate the architectural superiority of the proposed INRA-Transformer, a comparative benchmark experiment was conducted using the standard Vision Transformer (ViT) baseline under identical dataset conditions, spatial resolution, and hyperparameter configurations. The empirical comparison highlights the significant performance gain achieved by embedding multi-scale convolutional priors into the Transformer framework.

Table 2 Performance comparison between the Standard ViT and the proposed INRA-Transformer model.				
Model / Metric	Overall Accuracy (OA) (%)	Average Accuracy (AA) (%)	Kappa Coefficient (K) (%)	Test Loss (%)
Standard ViT	94.22	91.88	91.86	16.12
INRA-Transformer (Proposed)	97.58	96.03	96.60	-
Improvement (Delta)	+ 3.36%	+ 4.15%	+ 4.74%	Significant Reduction

Table 3
Detailed class-wise performance breakdown (Precision, Recall, F1-score).

Class	Model	Precision	Recall	F1-score
Low deformation	Standard ViT	0.92	0.93	0.92
	INRA-Transformer	0.97	0.95	0.96
Very high deformation	Standard ViT	0.90	0.95	0.92
	INRA-Transformer	0.98	0.96	0.97
Stabil	Standard ViT	0.96	0.94	0.95
	INRA-Transformer	0.97	0.98	0.98
Moderate deformation	Standard ViT	0.74	0.82	0.78
	INRA-Transformer	0.89	0.91	0.90
High deformation	Standard ViT	0.98	0.96	0.97
	INRA-Transformer	0.99	0.99	0.99

As presented in Table 2 and Table 3, the proposed INRA-Transformer demonstrates a clear superiority across all evaluation metrics compared to the standard ViT architecture:

Overall Classification Efficiency: The INRA-Transformer achieves a 3.36% increase in Overall Accuracy (OA) (rising from 94.22% to 97.58%) and a 4.74% boost in Kappa accuracy (from 91.86% to 96.60%). This indicates that the hybrid framework substantially reduces random misclassifications and provides stronger agreement with ground truth data.

Mitigating the Weakness in Transitional Classes: Standard ViTs often struggle with complex, continuous spatial phase boundaries due to rigid linear patch embeddings. This limitation is evident in the Moderate deformation class, where standard ViT yields a low F1-score of 0.78 (Precision: 0.74, Recall: 0.82). In contrast, the INRA-Transformer elevates the F1-score for Moderate deformation to 0.90 (Precision: 0.89, Recall: 0.91), achieving a massive + 12% performance jump.

Feature Extraction Superiority: The standard ViT exhibits a high Test Loss of 16.12%, indicating higher uncertainty during prediction. The integration of the Inception multi-scale stem in our model captures subtle phase gradients across 1x1, 3x3, 5x5 scales, effectively providing the local inductive bias that standard ViTs lack.

Rupture Core Delineation: In high-intensity displacement zones (High deformation and Very high deformation), the INRA-Transformer reaches near-perfect precision values (0.99 and 0.98, respectively), outperforming the standard ViT across all fringe boundaries.

These comparative results conclusively confirm that injecting multi-scale convolutional priors and depthwise-separable Feed-Forward Networks into Vision Transformers is essential for accurate DInSAR phase fringe interpretation and seismic hazard mapping.

5. Conclusion

In this study, a hybrid deep learning framework, named INRA-Transformer, was developed to classify post-seismic surface deformation patterns from Sentinel-1A DInSAR interferograms covering the 6.1 Mw Sındırgı earthquake event in Balıkesir, Turkey. The primary objective was to overcome the limitations of standard deep learning models and traditional Vision Transformers in accurately resolving fine-grained interferometric phase gradients amidst decorrelation noise and atmospheric phase screens. The proposed architecture addresses these challenges by introducing two key methodological innovations:

Multi-Scale Feature Extraction: An Inception-style convolutional stem replaces the rigid linear patch embedding of standard ViTs, capturing both high-frequency micro-scale phase transitions and broad tectonic deformation trends simultaneously.

Local Inductive Bias Preservation: A Depthwise-Convolutional Feed-Forward Network (DW-FFN) is embedded within the Transformer encoder stack, maintaining spatial continuity across phase fringe boundaries while leveraging global Multi-Head Self-Attention (MHSA).

To evaluate the effectiveness of the proposed hybrid approach, six comprehensive evaluation metrics were employed: class-specific accuracy, precision, recall, F1-score, Overall Accuracy (OA), Average Accuracy (AA), and the Kappa coefficient (K). Experimental results on a validation dataset of 50,854 pixels demonstrated that the INRA-Transformer achieved exceptional performance, reaching an Overall Accuracy of 97.58%, an Average Accuracy of 96.03%, and a Kappa coefficient of 96.60%.

When analyzing individual displacement classes, the model exhibited its highest classification success in the High deformation zone with a 99.05% accuracy and 0.99 F1-score, successfully delineating the primary seismic rupture core. Comparative benchmark experiments against the standard Vision Transformer (ViT) further underscored the superiority of our framework: the INRA-Transformer achieved a + 3.36% increase in OA (from 94.22% to 97.58%), a + 4.74% boost in Kappa accuracy, and a + 12% F1-score improvement in the challenging Moderate deformation class (jumping from 0.78 to 0.90). Furthermore, Explainable AI (XAI) attention map visualizations verified that the network dynamically prioritizes geophysically consistent phase gradients while effectively filtering out random atmospheric noise.

In conclusion, these findings demonstrate that integrating multi-scale convolutional priors with Vision Transformer architectures provides a highly reliable, accurate, and interpretable solution for automated DInSAR analysis. In future studies, we plan to extend the INRA-Transformer architecture to multi-temporal SAR time-series datasets and evaluate its generalization capability across different satellite frequencies (e.g., L-band and X-band) and complex tectonic environments.

Author's contributions Í. A.: Investigation, Software, Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Resources, Data collection, Analysis, Writing – review & editing, Supervision, Validation.

Declarations

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for Publication

Informed consent was obtained from all individual participants included in the study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not for-profit sectors.

Author Contribution

Author's contributions Í. A.: Investigation, Software, Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Resources, Data collection, Analysis, Writing – review & editing, Supervision, Validation. Data availability Data will be made available on request. Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not for-profit sectors. Declarations Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Ethics approval and consent to participate This article does not contain any studies with human participants or animals performed by any of the authors. Consent for Publication Informed consent was obtained from all individual participants included in the study.

Data availability

References

1. Albino F, Biggs J (2021) Magmatic processes in the East African Rift system: Insights from a 2015–2020 Sentinel-1 InSAR survey. *Geochem Geophys Geosyst*, 22(3), Article e2020GC009488.
2. Aslan İ, Polat N (2024) Deep learning-based classification of mature and immature lavender plants using UAV orthophotos and a hybrid CNN approach. *Earth Sci Inf* 17(2):1713–1727
3. Bakon M, Czikhardt R, Papco J, Barlak J, Rovnak M, Adamisin P, Perissin D (2020) remotIO: A Sentinel-1 multi-temporal InSAR infrastructure monitoring service with automatic updates and data mining capabilities. *Remote Sens*, 12(11), Article 1892.
4. Daramola OE, Epuh EE, Okolie CJ, Arungwa ID, Apeh OI, Fatoyinbo A, Tenzer R (2025) Detection of crustal deformation using a gravimetric and multitemporal interferometric SAR (MT-InSAR) approach. *Pure appl Geophys* 182(5):2111–2134
5. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Dosovitskiy A, Houlsby N (2020) An image is worth 16x16 words: Transformers for image recognition at scale (arXiv Preprint arXiv:2010.11929). *arXiv*
6. He J, Chen J, Xu H, Yu Y (2024) SonarNet: Hybrid CNN-transformer-HOG framework and multifeature fusion mechanism for forward-looking sonar image segmentation. *IEEE Trans Geosci Remote Sens* 62:1–17
7. Ji C, Wang L, Qin J, Liu L, Han Y, Wang Z (2024) MSGformer: A multi-scale grid transformer network for 12-lead ECG arrhythmia detection. *Biomed Signal Process Control*, 87, Article 105499.
8. Kumar N, Devanaboyina VR (2026) A review on marine atmospheric boundary layer roll detection and characterization using SAR satellite imagery data. *J Indian Soc Remote Sens*, 1–18
9. Li J, Yu Z, Yu L, Cheng P, Chen J, Chi C (2023) A comprehensive survey on SAR ATR in deep-learning era. *Remote Sens*, 15(5), Article 1454.
10. Liu X, Zhang Y, Shan X, Wang Z, Gong W, Zhang G (2025) Deep learning for automatic detection of volcanic and earthquake-related InSAR deformation. *Remote Sens*, 17(4), Article 686.
11. Ni K, Yuan C, Zheng Z, Zhang B, Wang P (2024) MPT-SFANet: Multiorder pooling transformer-based semantic feature aggregation network for SAR image classification. *IEEE Trans Aerosp Electron Syst* 60(4):4923–4938
12. Ross N, Milillo P, Dini L (2024) Automated grounding line delineation using deep learning and phase gradient-based approaches on COSMO-SkyMed DInSAR data. *Remote Sens Environ*, 315, Article 114429.
13. Shafapourtehrany M, Batur M, Shabani F, Pradhan B, Kalantar B, Özener H (2023) A comprehensive review of geospatial technology applications in earthquake preparedness, emergency management, and damage assessment. *Remote Sensing*, 15(7), Article 1939

14. Sun X, Wang P, Lu W, Zhu Z, Lu X, He Q, Li J, Fu K (2022) RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans Geosci Remote Sens* 61:1–22
15. Wei Y, Qiu H, Liu Z, Huangfu W, Zhu Y, Liu Y, Kamp U (2024) Refined and dynamic susceptibility assessment of landslides using InSAR and machine learning models. *Geosci Front*, 15(6), Article 101890.
16. Wassie Y, Milillo P (2025) Interferometric synthetic aperture radar multitemporal deformation monitoring: A review of machine learning techniques. *IEEE Geoscience Remote Sens Magazine* 13(2):220–243
17. Xie Y (2025) Deep learning in earthquake engineering: A comprehensive review. *ASCE OPEN: Multidisciplinary J Civil Eng*, 3(1), Article 03125001.
18. Zhang Q, Li Y, Zhang J, Luo Y (2019) InSAR technique applied to the monitoring of the Qinghai–Tibet Railway. *Nat Hazards Earth Syst Sci* 19(10):2229–2240
19. Zhao X, Pan S, Sun Z, Guo H, Zhang L, Feng K (2021) Advances of satellite remote sensing technology in earthquake prediction. *Nat Hazards Rev*, 22(1), Article 03120001.
20. Zhou L, Gao J, Chen H (2025) Seismic facies classification based on multilevel wavelet transform and multiresolution transformer. *IEEE Trans Geosci Remote Sens* 63:1–12
21. Zhou L, Yu H, Lan Y (2021) Artificial intelligence in interferometric synthetic aperture radar phase unwrapping: A review. *IEEE Geoscience Remote Sens Magazine* 9(2):10–28
22. Zhu XX, Montazeri S, Ali M, Hua Y, Wang Y, Mou L, Bamler R (2021) Deep learning meets SAR: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience Remote Sens Magazine* 9(4):143–172
23. Zhu XX, Tuia D, Mou L, Xia GS, Zhang L, Xu F, Fraundorfer F (2017) Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience Remote Sens Magazine* 5(4):8–36

Figures

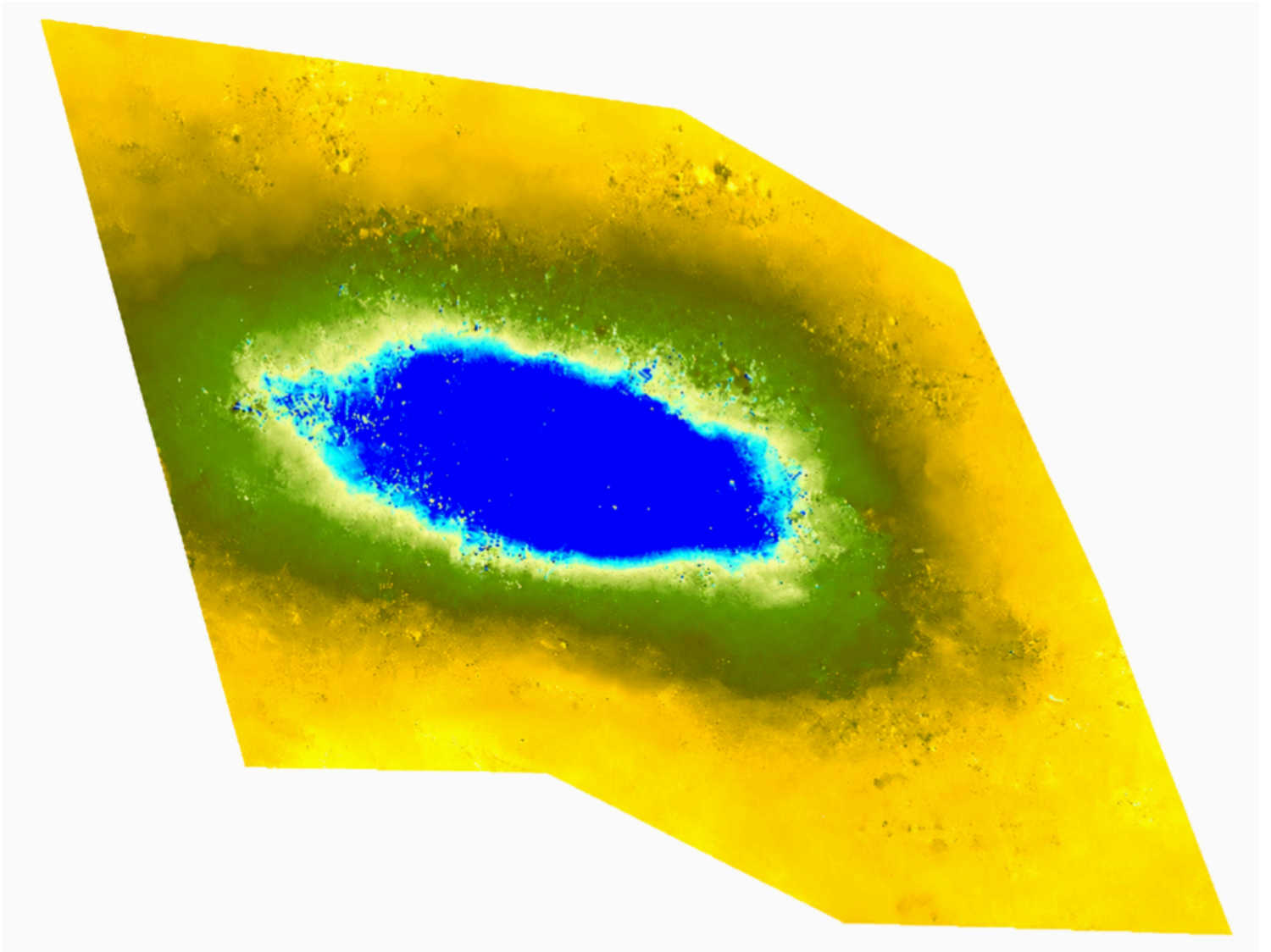


Figure 1

Location of the Sındırgı earthquake epicenter and regional fault systems.

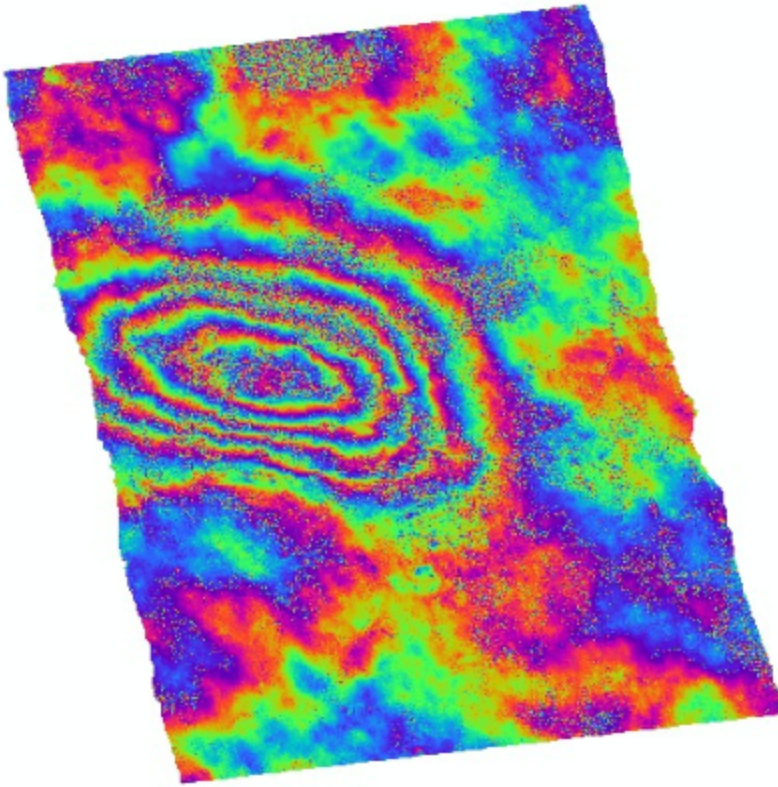


Figure 2

Differential interferogram of the Sındırgı earthquake showing wrapped phase fringes.

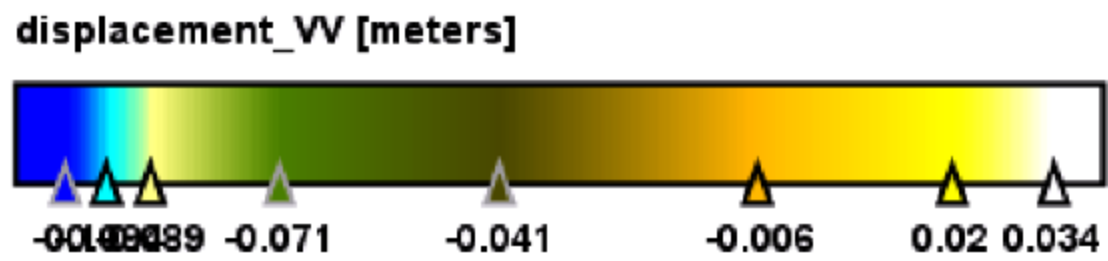


Figure 3

Technical legends for Wrapped phase and LOS displacement in meters, used as the basis for ground truth labeling.

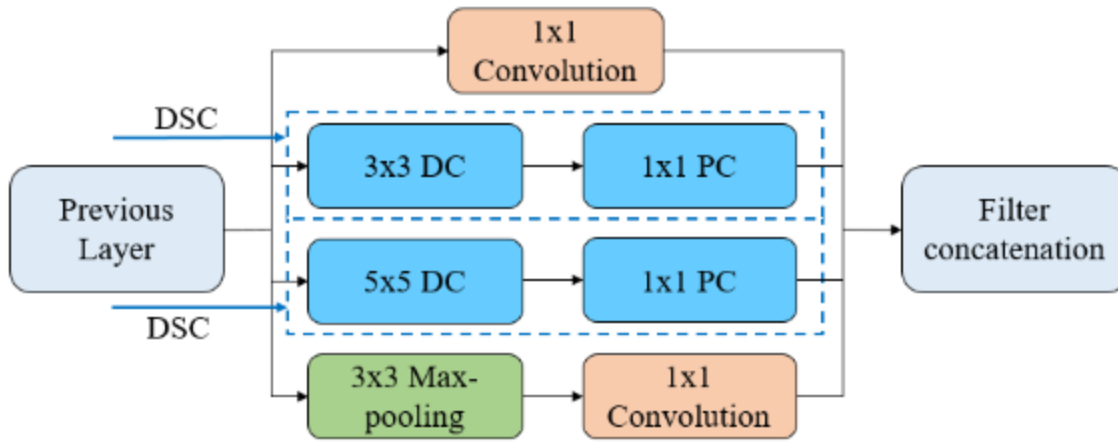


Figure 4

Structure of the Inception depthwise separable convolution block

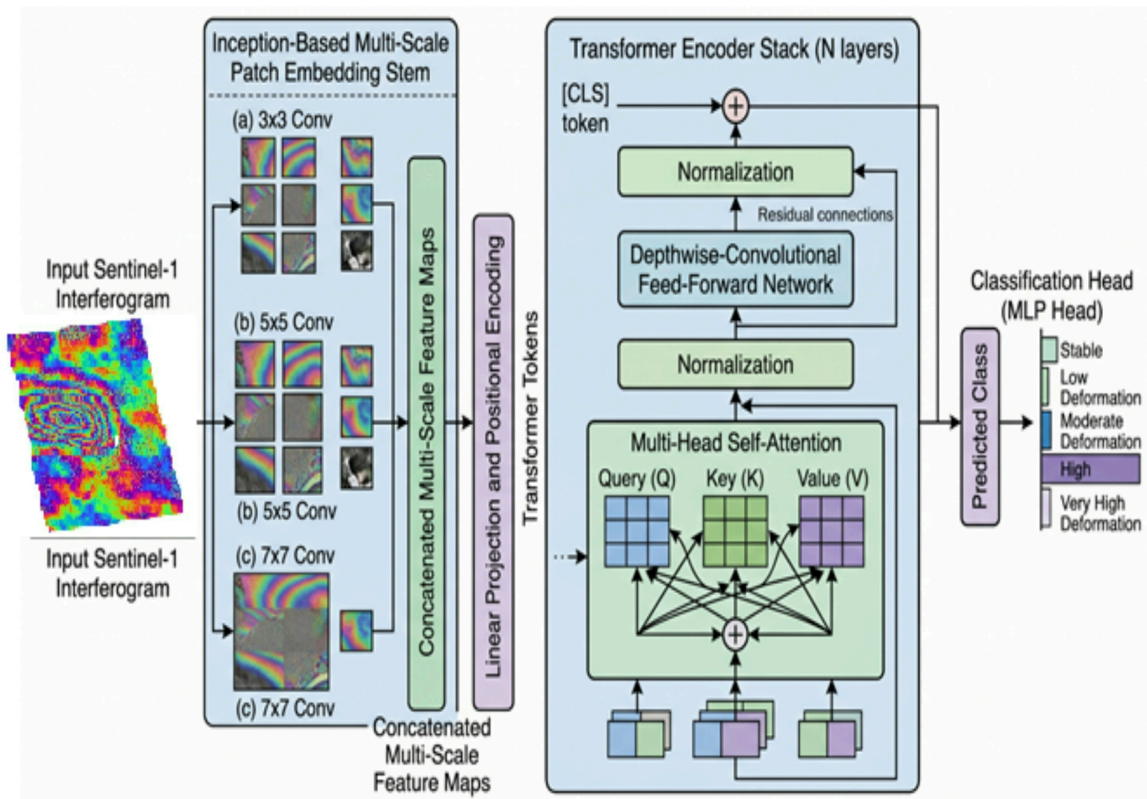


Figure 5

Proposed INRA-Transformer Architecture

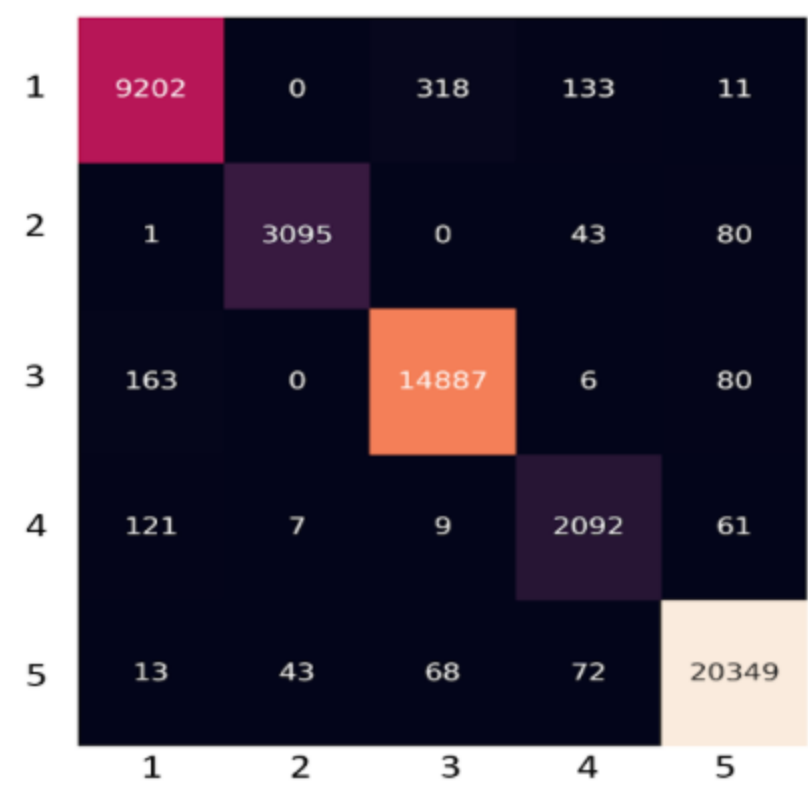
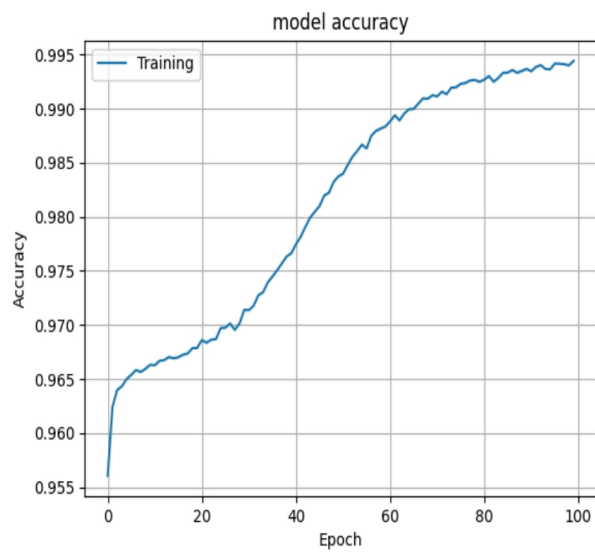
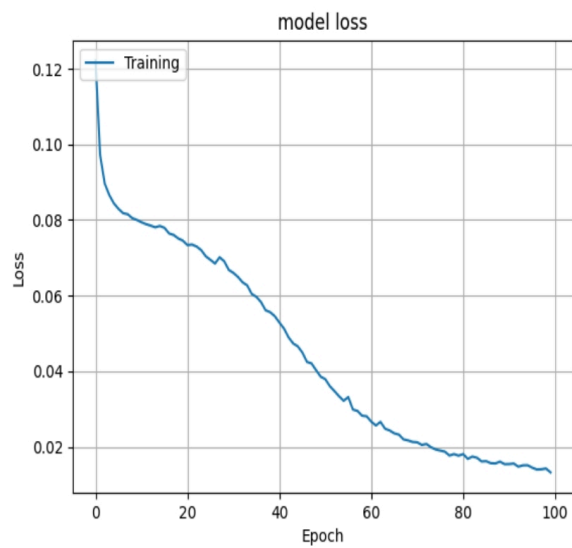


Figure 6

Confusion matrix for the validation dataset (50,854 pixels) generated by the INRA-Transformer model. Rows represent the actual ground-truth classes, and columns represent the predicted classes. Class labels correspond to: (1) Low deformation, (2) Very high deformation, (3) Stable, (4) Moderate deformation, and (5) High deformation.



(a)



(b)

Figure 7

(a) Model accuracy, (b) Model loss curves of the proposed INRA-Transformer method over 100 training epochs.

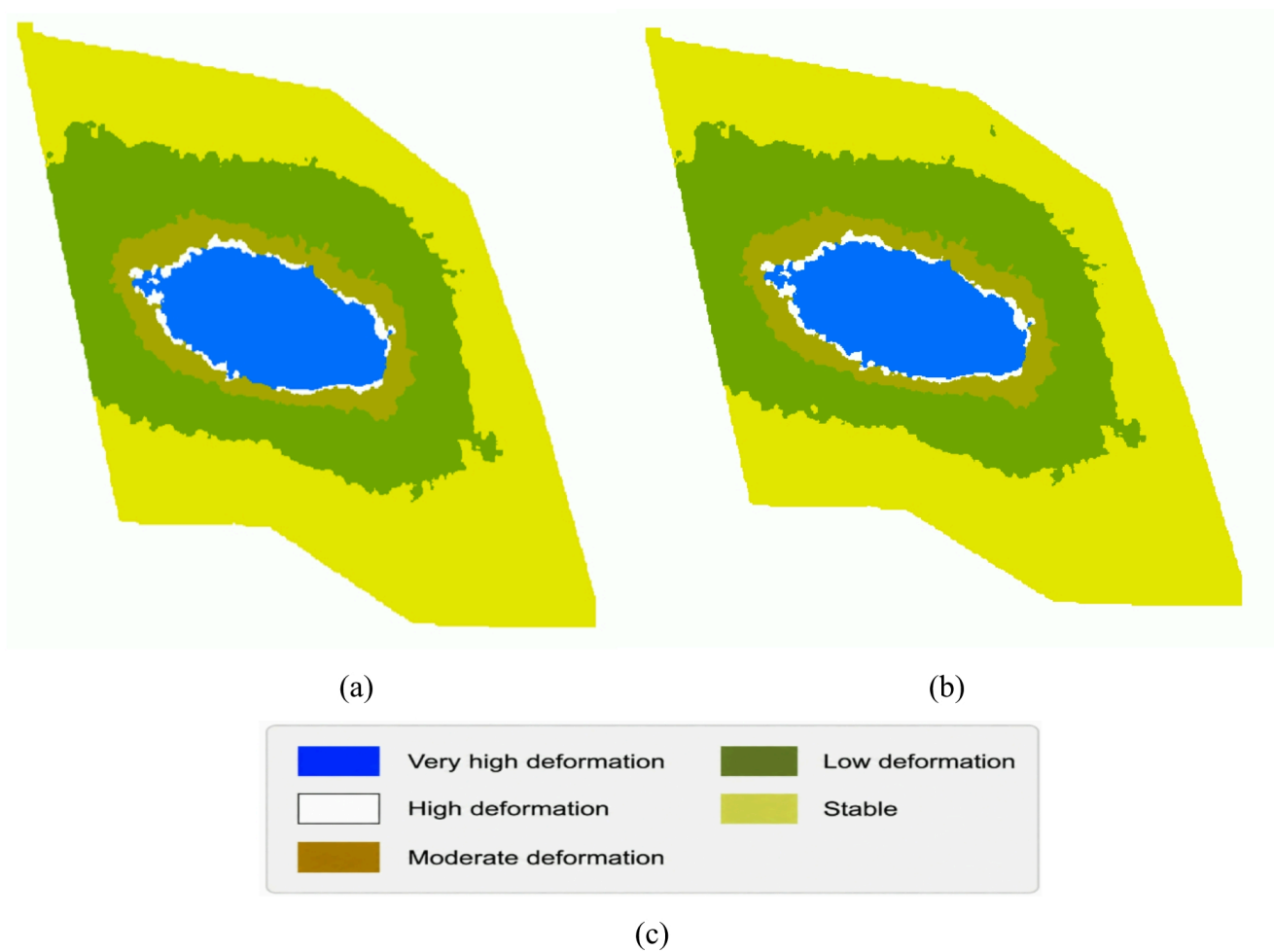


Figure 8

Spatial classification results over the Sındırgı earthquake area: (a) Ground truth map, (b) INRA-Transformer prediction map, and (c) Deformation class legend.

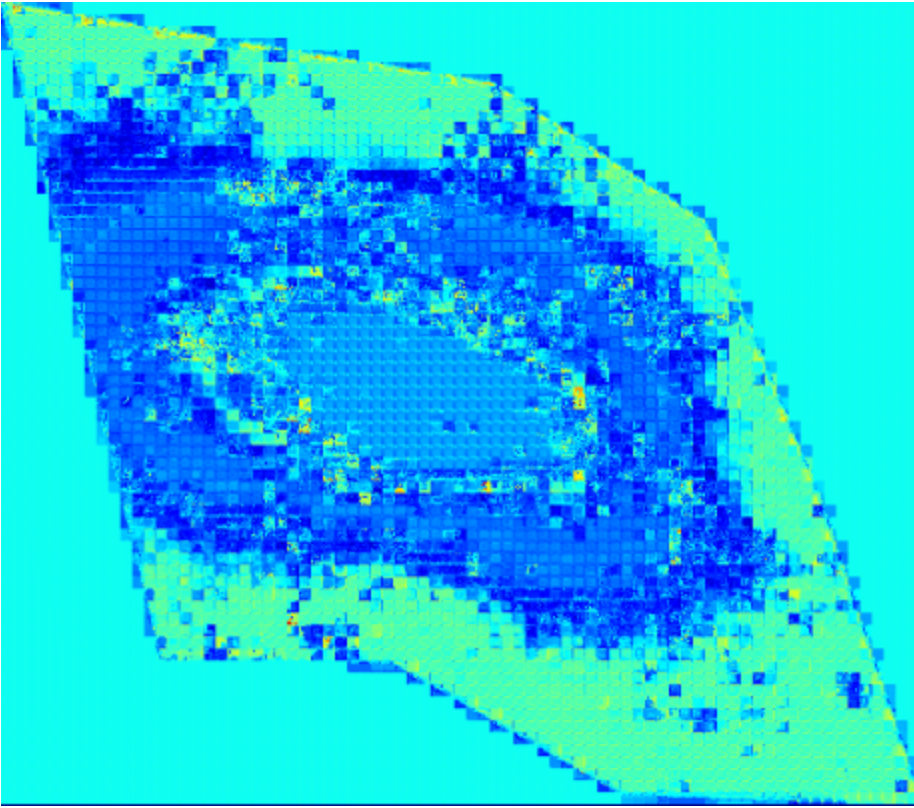


Figure 9

Attention map visualizations generated by the final self-attention layer of the INRA-Transformer model, highlighting spatial focus regions across the Sındırgı earthquake interferogram.