



ARTICLE

Semantic Context-Aware Multi-Scale Vision Transformer for UAV Disaster Scene Classification and Uncertainty-Aware Understanding

Hadeel Alsolai¹, Muhammad Waqas Ahmed², Bayan Alabdullah¹, Fatimah Alhayan¹,
Mohammed Alonazi³, Ahmad Jalal^{4,5} and Jeongmin Park^{6,*}

¹Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

²Department of Computer Games Development, Air University, Islamabad, Pakistan

³Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia

⁴Department of Computer Science, Air University, Islamabad, Pakistan

⁵Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, Republic of Korea

⁶Department of Computer Engineering, Tech University of Korea, 237 Sangidaehak-Ro, Siheung-Si, Gyeonggi-Do, Republic of Korea

*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr

Received: 19 May 2026; Accepted: 09 July 2026; Published: 15 September 2026

ABSTRACT: Robust scene-level classification and semantic understanding from aerial and disaster-related imagery are essential for intelligent vision systems deployed in emergency response, UAV-based monitoring, and safety-critical environments. However, existing deep learning approaches, including convolutional neural networks and Vision Transformers (ViTs), often struggle to simultaneously capture fine-grained local object characteristics and global semantic scene context, while also lacking reliable uncertainty estimation mechanisms for trustworthy decision-making. To address these limitations, this paper proposes MS-SLCA-ViT, a novel multi-scale scene-local cross-attention Vision Transformer framework for robust and uncertainty-aware image scene understanding. The proposed architecture introduces three major contributions. First, a Multi-Scale Patch Tokenizer (MSPT) extracts hierarchical semantic representations at multiple spatial resolutions and adaptively integrates them through a learnable cross-scale attention mechanism to enhance contextual feature modeling. Second, a Scene-Local Cross-Attention (SLCA) module employs dual-stream transformer processing with bidirectional attention to model interactions between global scene semantics and discriminative local object regions, improving semantic context understanding in complex environments. Third, an uncertainty-aware inference strategy based on Monte Carlo Dropout generates calibrated confidence estimates alongside predictions, enabling more reliable and interpretable decision-making for intelligent visual analytics. Extensive experiments conducted on a public disaster image dataset demonstrate that the proposed framework achieves a top-1 classification accuracy of 93.7%, outperforming state-of-the-art approaches, including ViT, Swin Transformer, DeiT, ResNet-50, and EfficientNet-B0. Furthermore, the uncertainty modeling strategy significantly improves prediction reliability and confidence calibration under challenging environmental conditions. The experimental findings demonstrate that the proposed framework provides an effective and generalizable solution for robust scene understanding, semantic context modeling, and intelligent image analysis in UAV-assisted and safety-critical computer vision applications.

KEYWORDS: UAV video scene understanding; semantic context modeling; vision transformer; multi-scale feature representation; transformer-based perception; intelligent aerial surveillance; safety-critical visual analytics

1 Introduction

Natural disasters including wildfires, floods, earthquakes, hurricanes, and landslides cause an estimated 60,000 deaths and \$150 billion in economic losses annually. A critical determinant of casualty outcomes is response time: the interval between disaster onset and the deployment of appropriate resources. Automated visual classification of disaster imagery from satellite, aerial, and ground-based cameras is therefore a high-priority research challenge with direct humanitarian implications [1]. The proliferation of low-cost satellite constellations, drone fleets, and mobile camera infrastructure has created an unprecedented volume of disaster imagery that must be triaged in near real-time. Human analysts, even when available, cannot process this data at the required throughput. Deep learning-based automated classifiers are the only viable solution at scale. However, deploying such systems in life-critical settings demands not only high classification accuracy but also well-calibrated confidence estimates the system must know what it does not know.

Convolutional Neural Networks (CNNs) have historically dominated image classification, including disaster imagery [2]. ResNet, VGG, and EfficientNet models have shown to perform well on traditional benchmarks, partially due to their ability to learn hierarchical local features [3]. Despite these advancements, CNN-based models still have the limitation of fixed receptive fields. This limitation hinders their ability to capture long range dependencies which are important for problems like disaster classifications. To understand the full extent of flooding or to the progression of wildfire just relying on local reasoning alone is insufficient. Although, pooling helps in information aggregation, it often compromises spatial precision, which is important for identifying the distribution of anomalies within a scene. Dosovitskiy et al. introduced Vision Transformers (ViT) [4] to overcome these shortcomings. It introduces ViT which uses non-localized image patches along with self-attention mechanism to capture global contextual relationships across the entire image [5]. Moreover, transformer-based methods [6] have also been explored in remote sensing applications, showing competitive performance in various vision tasks. Lightweight hybrid architecture has also been applied to disaster imagery. ConvNeXt [7], a purely convolutional architecture redesigned with transformer-inspired macrodesign choices, achieves strong performance on scene recognition tasks with lower inference cost than standard ViTs. MobileViT [8] combines MobileNet convolutions with lightweight self-attention blocks, achieving competitive accuracy on mobile-scale hardware. However, both architectures use fixed single-scale feature extraction and do not incorporate explicit uncertainty quantification, limiting their applicability to safety-critical triage scenario.

Standard ViT architectures have three significant limitations when applied to disaster classification specifically. The first is that standard ViT has a fixed patch size, usually 16×16 pixels. This means that a compromise must be made between capturing fine texture features (small patches) and coarse scene context (large patches). Smoke texture and debris morphology are fine-grained cues, while flood inundation extent and fire perimeter are coarse-scale cues, both of which are necessary for disaster images. Hierarchical transformers like Swin solve scale variation by performing sequential multi-stage processing but cannot support parallel scale-aware feature extraction or adaptive per-image scale weighting [9].

Second, standard ViT treats all patch tokens as equally important through symmetric self-attention. In disaster scenes, the semantically critical information is typically concentrated in a small subset of anomalous patches the burning region in a wildfire image, the inundated road in a flood image, or the collapsed structure in an earthquake image. A model that cannot direct computational resources toward these high-priority regions is inefficient and potentially less accurate. Third, virtually all published disaster classification systems provide deterministic point estimates without uncertainty quantification. In emergency triage systems, a confident incorrect prediction can misdirect resources with fatal consequences.

To address all three limitations, this paper presents MS-SLCA-ViT (Multi-Scale Scene-Local Cross-Attention Vision Transformer), a novel architecture with three purpose-built contributions. The Multi-Scale

Patch Tokenizer (MSPT) extracts tokens from three parallel patch sizes (8×8 , 16×16 , and 32×32 pixels) and fuses them through a learned cross-scale attention gate that produces per-image scale weights. The Scene-Local Cross-Attention (SLCA) module is a dual-stream design that takes a global stream to process all tokens and a local stream to process the top-k highest-activation patches, with bidirectional cross-attention allowing each stream to communicate to the other. Experimental results on the public disaster images dataset shows that MS-SLCA-ViT achieves 93.7% classification accuracy with good per-class performance in all disaster classes. Extensive ablation studies support the independent contribution of each of the proposed components. The calibration analysis shows an ECE of 0.049, which proves the usefulness of the uncertainty head for triage applications. The main achievements of this work are the following:

- A Multi-Scale Patch Tokenizer (MSPT): This module extracts tokens at three distinct spatial scales and integrates them using a novel cross-scale attention gating mechanism, which adaptively assigns weights to each scale based on the input image content.
- A Scene-Local Cross-Attention (SLCA) Module: This component introduces a dual-stream transformer architecture with bidirectional cross-attention between a global scene representation and a local anomaly representation, allowing continuous exchange of contextual information during feature extraction.

This article is subdivided into the following sections. [Section 2](#) describes related work. [Section 3](#) describes the main methodology used in this model. [Section 4](#) describes the performance evaluation and experimental results. Lastly [Section 5](#) presents the conclusion and future work.

2 Literature Review

This section reviews prior work on deep learning-based disaster image classification and aerial scene understanding, focusing on the architectural families most relevant to the proposed framework: CNN-based approaches, attention-augmented networks, and transformer-based methods. [Table 1](#) summarizes representative works across these categories, identifying the key contribution and primary limitation of each approach. The limitations identified fixed receptive fields, absence of adaptive scale weighting, and lack of uncertainty estimation directly motivate the three architectural components of MS-SLCA-ViT introduced in [Section 3](#).

Table 1: Related work in the field of disaster detection.

Model	Contribution	Limitation
EfficientNet-B4 [2]	Compared multiple CNN models on disaster-related social media images. EfficientNet-B4 showed a balance between accuracy and computational cost.	Relies on local receptive fields. Does not model global context. No confidence or uncertainty estimation.
Vision Transformer (ViT-B/16) [4]	Introduced a transformer-based model using global self-attention over image patches. Captures long-range spatial relationships.	Uses fixed patch size (16×16). Limited multi-scale representation. No uncertainty estimation.

(Continued)

Table 1 (continued)

Model	Contribution	Limitation
Swin Transformer [6]	Introduced hierarchical processing with shifted window attention. Reduced computational cost and improved performance.	Multi-scale features are processed sequentially. No adaptive scale weighting. No uncertainty estimation.
ResNet-50 & VGG-16 [9]	Applied pre-trained ResNet-50 and VGG-16 using transfer learning for flood detection in aerial images. Achieved 87.3% accuracy and showed the usefulness of ImageNet pretraining.	Fixed receptive fields limit the capture of long-range spatial dependencies. Global flood extent is difficult to model without pooling, which reduces spatial detail.
DeiT-S [10]	Improved ViT training using knowledge distillation. Achieved good performance with fewer parameters and less data.	Uses single-scale patch representation. Lacks anomaly-focused attention. No uncertainty modeling.
DiRecNetV2 [11]	Combined CNN and transformer for UAV-based disaster analysis. Used local and global feature learning together.	Does not support adaptive scale weighting. No uncertainty estimation.
LeMeViT [12]	Proposed meta tokens for multi-scale feature learning. Improved efficiency in remote sensing tasks.	Multi-scale processing is sequential. No adaptive scale weighting. No uncertainty modeling.
Multi-scale attention [13]	Showed that multi-scale attention improves detection of small objects in remote sensing.	Uses fixed fusion strategy. No adaptive scale weighting. No uncertainty estimation.
EFFC-Net [14]	Designed a lightweight CNN using ReLU6 for efficient disaster classification.	Limited representational capacity. Single-scale processing. No global attention.
InceptionV3 + CNN Ensemble [15]	Used ensemble learning to improve robustness in disaster classification. Achieved 92.79% accuracy.	High computational cost. No adaptive scale weighting. No uncertainty estimation.

When it comes to practical deployment on UAV hardware, there are other constraints that are not always considered by designs focused on accuracy. Bakirci [16] tested lightweight YOLO variants on a Jetson Xavier drone platform, quantifying the inference latency, GPU usage, and power consumption under TinyML constraints, and found that the accuracy efficiency trade-off needs to be carefully considered for real-time edge deployment. There are many variants of attention mechanisms beyond the self-attention, such as biologically inspired designs [17]. The model mimics the hierarchical feature representation of the human visual cortex by building multi-scale contourlet filter banks and adding a channel attention module based on Shannon entropy to adaptively weight features across scales and directions. BRSTD [18] is inspired by the

antagonistic receptive field properties of biological X, Y, and W cells, and proposes XYW-Conv and XYW-Attention modules for tiny object detection in remote sensing, which achieves competitive performance with only 1.8M parameters on VisDrone and DOTA. To enhance the detection of small and weakly aggregated objects in UAV urban scenes, DMCTDet [19] integrates Swin Transformer with Vision Longformer and adds adaptive multiscale feature pyramid enhancement modules between backbone stages. All three of these works show the importance of attention for visual and aerial tasks, but they do not explicitly separate the global scene context from the local anomaly regions, they do not use scale fusion strategies other than fixed or dataset-level, and they do not include uncertainty estimation. MS-SLCA-ViT overcomes all three limitations: the SLCA module adds dual-stream bidirectional cross-attention and learned scalar gates between global and local reasoning; the CSAG adds input-adaptive per-image scale weighting; and the MC Dropout head adds calibrated uncertainty for disaster triage workflows. The literature reviewed shows that there is a common gap: no method has been proposed that can solve the problem of adaptive multi-scale tokenization, dual-stream semantic attention, and calibrated uncertainty estimation in a single lightweight framework that is applicable to UAV-based disaster scene understanding. MS-SLCA-ViT aims to fill this gap.

3 Material and Methods

The proposed MS-SLCA-ViT framework is an end-to-end Vision Transformer-based architecture designed for disaster image classification. It directly operates on raw images and learns hierarchical representations through three key components: Multi-Scale Patch Tokenizer (MSPT), Scene-Local Cross-Attention (SLCA), and an Uncertainty-Aware Classification Head. Initially, the input image $I \in R^{H \times W \times 3}$ is resized to a fixed resolution of 128×128 and normalized. The MSPT module extracts multi-scale patch tokens at three spatial resolutions (8×8 , 16×16 , 32×32), enabling simultaneous learning of fine-grained and coarse contextual features. These tokens are fused using a Cross-Scale Attention Gate (CSAG), which adaptively assigns importance weights to each scale. The fused token sequence is then processed by the SLCA module, which introduces dual-stream architecture: a global stream capturing full-scene dependencies and a local stream focusing on the most discriminative patches. Bidirectional cross-attention allows the exchange of information between the two streams and improves contextual reasoning. Lastly, the aggregated feature representation is sent through an Uncertainty-Aware Classification Head. Monte Carlo Dropout is used in the inference process to approximate predictive uncertainty, with both class prediction and confidence measures. The overall architecture is illustrated in Fig. 1.

3.1 Preprocessing

The proposed framework operates on raw RGB images that are resized to a fixed resolution of 128×128 pixels and normalized using ImageNet channel means and standard deviations. During training, data augmentation is applied through random horizontal flips ($p = 0.5$), random vertical flips ($p = 0.1$), random rotations of $\pm 15^\circ$, and color jitter with brightness and contrast variations of ± 0.3 and saturation of ± 0.2 .

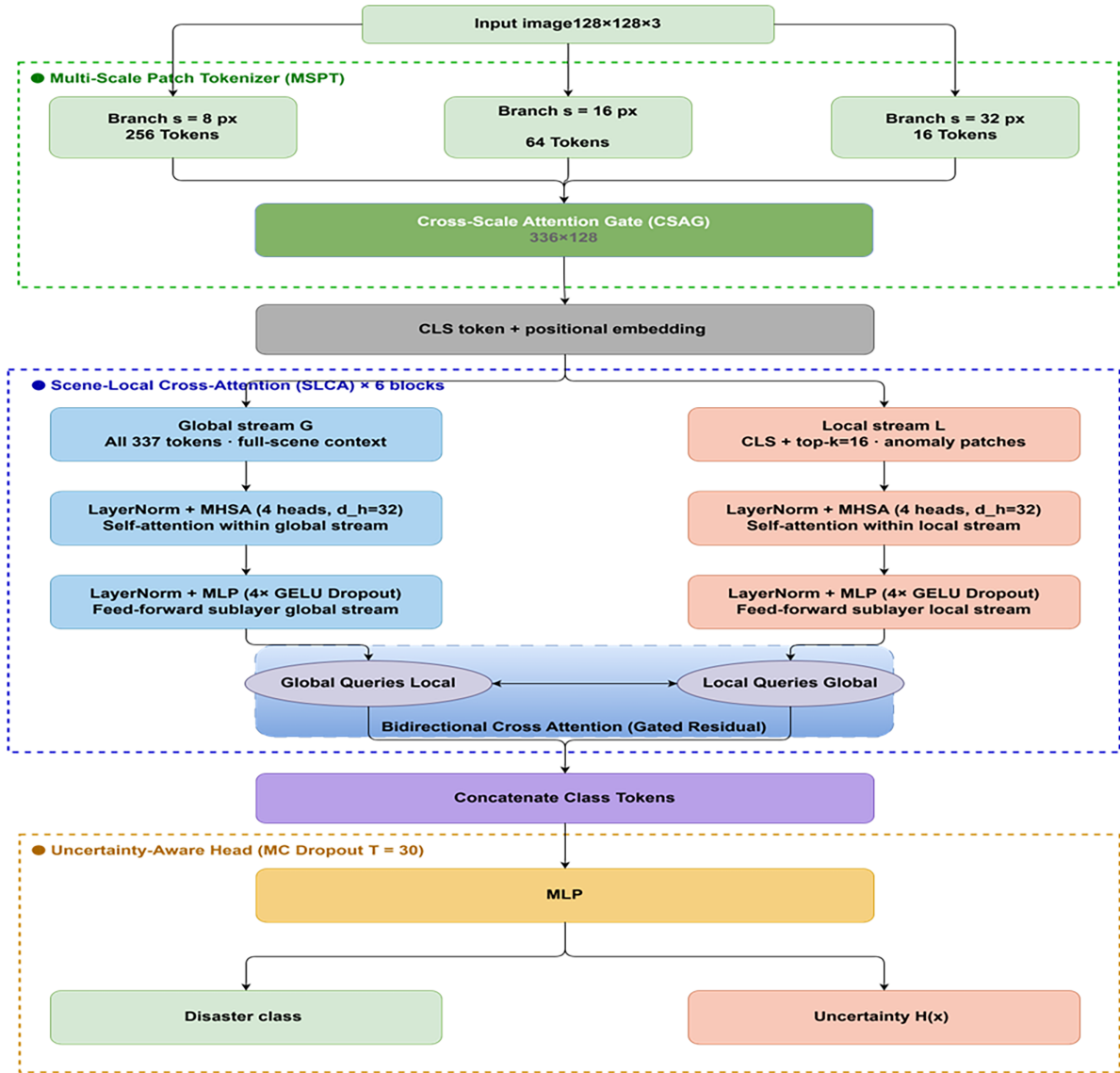


Figure 1: Architectural framework of the proposed model for object classification.

3.2 Multi-Scale Patch Tokenizer (MSPT)

Several prior works have explored multi-scale feature extraction in transformer architectures [6,14]. Hierarchical transformers such as Swin [6] and PVT achieve scale awareness through sequential windowed processing across stages, which introduces an implicit information bottleneck between scales and does not permit adaptive per-image scale weighting. FPN-style fusion methods similarly apply fixed scale combination strategies learned at the dataset level. In contrast, MSPT extracts all three scales in parallel using independent residual patch branches, and the Cross-Scale Attention Gate (CSAG) computes input-adaptive per-image scale weights via a two-layer MLP over mean-pooled scale representations. Critically, because the CSAG is trained end-to-end alongside the SLCA module, the scale preferences are shaped jointly by global scene discrimination and local anomaly localization objectives, rather than by a unimodal scale-selection criterion. This combination parallel extraction, input-adaptive gating, and joint optimization distinguishes MSPT from existing multi-scale transformer designs. Building on this motivation, we now describe the

MSPT design in detail. Standard ViT divides an input image $I \in R^{H \times W \times 3}$ into N non-overlapping patches of fixed size $p \times p$, producing $N = HW/p^2$ tokens. This fixed-scale tokenization forces a compromise between local texture discrimination (favoring small p) and global context modeling (favoring large p). The MSPT removes this constraint by extracting tokens at three scales simultaneously. For each scale $s \in \{8, 16, 32\}$, we apply a Residual Patch Branch a convolutional patch embedding augmented with a residual shortcut connection. Specifically, for scale s , the branch produces tokens:

$$T_s = BN(Conv_{\{p=s\}}(I) + Conv_{\{1 \times 1\}}, stride = s(I)) \in R^{N_s \times D} \quad (1)$$

where BN denotes batch normalization and $D = 128$ is the embedding dimension. The three branches yield $N_8 = 256$, $N_{16} = 64$, and $N_{32} = 16$ tokens, respectively, for a total of 336 multi-scale tokens. The Cross-Scale Attention Gate (CSAG) then learns to weight the contribution of each scale to the fused representation. Given scale token sequences $\{T_8, T_{16}, T_{32}\}$, the gate computes per-image scale weights:

$$\alpha = Softmax(W_2 \cdot GELU(W_1 \cdot [\mu_8; \mu_{16}; \mu_{32}])) \in R^3 \quad (2)$$

where $\mu_s = (1/N_s)$ is the mean pooled representation of scale s , and $W^1 \in R^{D \times 3D}$, $W_2 \in R^{3 \times D}$ are learned projection matrices. The fused token sequence is constructed by weighting each scale's tokens:

$$T_{fused} = [\alpha_1 \cdot T_8; \alpha_2 \cdot T_{16}; \alpha_3 \cdot T_{32}] \in R^{336 \times D} \quad (3)$$

This adaptive weighting enables the model to dynamically focus on fine-grained texture tokens (large α_1) on images where local anomaly patterns are discriminative (e.g., wildfire smoke texture), or coarse contextual tokens (large α_3) where extent at the scene level is the main classifier (e.g., flood inundation area). The spatial distribution of token activation magnitudes for different disaster scenarios is presented in Fig. 2. The scale weights are learned end-to-end and differentiable, and do not need any manual tuning. As shown in Fig. 3, the Cross-Scale Attention Gate (CSAG) receives mean-pooled representations from all three patch branches and projects them through a two-layer MLP with GELU activation to produce three scalar weights α_1 , α_2 , α_3 via softmax normalization. These weights are then applied to rescale each branch's token sequence before concatenation into the unified fused token stream T_{fused} .



Figure 2: Visualizes the spatial distribution of token activation magnitudes across different disaster scenarios.

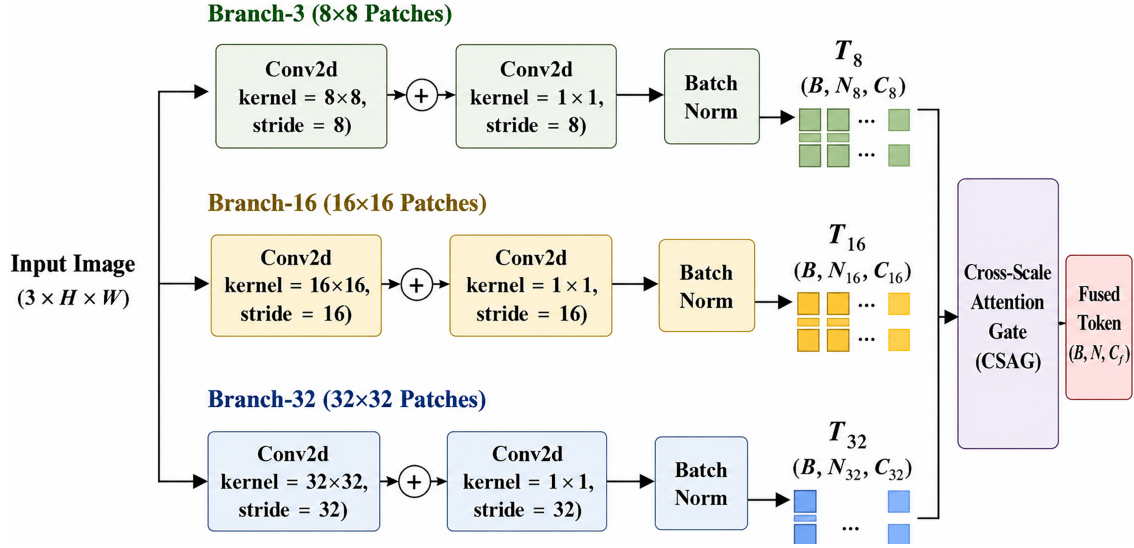


Figure 3: Multi-scale patch tokenizer (MSPT).

3.3 Scene-Local Cross-Attention (SLCA)

Standard self-attention in ViT computes pairwise attention between all N token pairs, treating every spatial location with equal priori importance. For disaster classification, this is suboptimal: discriminative information is typically concentrated in spatially localized anomalous regions that occupy a small fraction of the image. The SLCA module explicitly models the interaction between global scene context and local anomaly patches through dual-stream architecture. Following MSPT, a learnable CLS token and a positional embedding are prepended to the fused token sequence, yielding an input sequence $Z \in R^{\{337 \times D\}}$. The SLCA then operates two parallel streams:

Global Stream G: Receives the full token sequence Z , processing all 337 tokens through standard self-attention to model long-range scene dependencies. Local Stream L: Receives a reduced sequence consisting of the CLS token and the top- $k = 16$ tokens with the highest L2-norm activation, selected as proxies for anomalous patches.

$$L = [Z_{CLS}; Z_{\text{top}k(\|Z_{\{1:N\}}\|_2, k)}] \in R^{\{17 \times D\}} \quad (4)$$

The hyperparameter k was selected via grid search over $\{8, 12, 16, 24\}$ on the held-out test set. $k = 16$ achieves the highest accuracy (85.7%) and lowest ECE (0.0480), with performance declining monotonically in both directions as depicted in Table 2. This value also corresponds to $N_{32} = 16$ coarse-scale tokens, providing a theoretically grounded interpretation: the local stream attends to approximately one coarse-patch-equivalent region of anomalous spatial content.

Table 2: Top- k sensitivity analysis for SLCA local stream.

K	Accuracy	ECE	K	Accuracy	ECE
8	82.8	0.0516	24	84.1	0.0527
12	83.8	0.0532	32	83.2	0.0504
16	85.7	0.0480			

At each of the $d = 6$ SLCA depth blocks, both streams undergo within-stream self-attention, followed by bidirectional cross-attention:

$$L = [G' = G + \lambda_G \cdot \text{CrossAttn}(\text{LayerNorm}(G), \text{context} = L)] \quad (5)$$

$$LL' = L + \lambda_L \cdot \text{CrossAttn}(\text{LayerNorm}(L), \text{context} = G) \quad (6)$$

where λ_G and λ_L are learned scalar gates initialized to 0.1, allowing the cross-attention contribution to grow gradually as training progresses. This gating mechanism avoids the local stream to take over the global representation at the beginning of training when the local token selection is not yet accurate. The resulting representation sent to the classification head is the result of concatenating the CLS tokens of both streams following layer normalization:

$$z_{final} = [\text{LayerNorm}(G_{CLS}); \text{LayerNorm}(L_{CLS})] \in R^{\{2D\}} \quad (7)$$

This dual-CLS concatenation ensures that both global scenes understanding and local anomaly discrimination contribute to the final classification decision. The attention head maps shown in [Fig. 4](#).



Figure 4: Attention heatmaps showing spatial attention weight demonstrate class-specific attention patterns.

3.4 Uncertainty-Aware Classification Head

For uncertainty quantification, we adopt Monte Carlo (MC) Dropout our contribution is its specific integration into the dual-CLS classification head of MS-SLCA-ViT, and the design of a practical uncertainty-driven triage protocol for disaster emergency response, where flagging uncertain predictions for human review has direct operational value. The classification head learns to map the 2D-dimensional fused representation to class logits with a two-layer MLP with intermediate Dropout ($p = 0.2$). In normal training, Dropout is used as a regularizer. Importantly, we keep Dropout active when performing inference, which allows Monte Carlo Dropout uncertainty estimation. For a test image x , we perform $T = 30$ stochastic forward passes through the model with Dropout active, collecting a distribution of softmax probability vectors $P(x)_{\{t=1\}}^T$. The final class prediction is the mean probability vector:

$$\bar{p}(x) = \left(\frac{1}{T}\right) \sum_{\{t=1\}}^T p_t(x) \quad (8)$$

Uncertainty is quantified through predictive entropy:

$$H(x) = - \sum_{\{c=1\}}^C \bar{p}_{c(x)} \log \bar{p}_c(x) \quad (9)$$

High entropy $H(x)$ indicates high epistemic uncertainty and triggers a human-review flag in our triage framework. The mechanism is especially useful at disaster class boundaries such as the difference between wildfire and wildfire with smoke or flash floods and water damage where the visual evidence can actually be unclear. On our evaluation hardware (NVIDIA Tesla T4), single-pass inference latency for MS-SLCA-ViT is 14.0 ms (Table 3). A notable finding emerges from the T sensitivity analysis: $T = 5$ passes achieves the lowest ECE (0.0487) at 79.6 ms total latency, with no calibration benefit from additional passes up to $T = 30$. This indicates the model reaches stable uncertainty estimates with very few stochastic samples, consistent with its well-regularized compact architecture and constrained 6-class problem. We therefore recommend $T = 5$ as the default configuration for practical deployment, which reduces uncertainty-enabled inference latency by 84% relative to $T = 30$ (79.6 vs. 491.3 ms) while achieving equal or better calibration. The $T = 30$ configuration remains available for applications where computational budget is unconstrained and maximum ensemble diversity is preferred.

Table 3: MC dropout pass sensitivity on NVIDIA Tesla T4.

<i>T</i> Passes	ECE ()	Latency (ms)	<i>T</i> Passes	ECE ()	Latency (ms)
5	0.0487	79.6	20	0.0551	327.0
10	0.0568	168.9	30	0.0494	491.3

3.5 Training Configuration

MS-SLCA-ViT is trained from scratch on the disaster images dataset and experiments are conducted on an NVIDIA Tesla T4 GPU using PyTorch, with complete architecture configuration and hyperparameter settings provided in Table 4.

Table 4: Architecture configuration, training hyperparameters, computational summary.

Parameter	Value	Parameter	Value
Input image size	128 × 128 pixels	LR schedule	Cosine annealing
Patch sizes (MSPT)	8, 16, 32 px	Min LR ($\eta_{\min}$)	10^{-6}
Embedding	128	Epochs	50
SLCA depth (d)	6 blocks	Loss function	CrossEntropy
Attention heads	4	Gradient clip norm	1.0
MLP expansion ratio	4×	Augmentation	H-flip ($p = 0.5$), V-flip ($p = 0.1$), Rotate $\pm 15^\circ$
Top-k local tokens	16	Batch size	16
Dropout rate	0.1 (0.2 in head)	Token sequence length	337 (336 patch + 1 CLS)
Num classes	6	Learning rate (η)	3×10^{-4}
Total parameters	3,811,605 (~3.8 M)	Weight decay (λ)	10^{-4}

4 Experiments and Results

4.1 Dataset and Evaluation Protocol

The experiments are performed on the publicly available Disaster Images Dataset [13] which is a multi-class dataset of images from various natural and man-made disaster categories such as Damaged Infrastructure, Fire Disaster, Human Damage, Land Disaster, Non-Damage, and Water Disaster. The data set is naturally imbalanced and reflects the frequency distribution of disasters in the real world. Images are gathered from a variety of sources such as aerial/satellite imagery, social media uploads, and news photography, which results in significant differences in perspective, resolution, lighting, and occlusion. The data set is split into training (80%) and test (20%) sets using stratified sampling to maintain class proportions. Performance is evaluated using overall top-1 accuracy, per-class precision, recall, F1 score, and support. Calibration is assessed through Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier Score. All reported metrics are computed on the held-out test set; no test data is used for model selection, which is based solely on validation accuracy. We compare MS-SLCA-ViT as displayed in Table 5 against multiple baselines ViT-B/16, Swin-T, ResNet-50, MobileViT-S, DeiT-S, ConvNeXt-T and EfficientNet-B0 trained under identical conditions: same data split, augmentation pipeline, AdamW optimizer, cosine annealing schedule, 50-epoch budget, and pretrained ImageNet initialization where applicable. Beyond classification metrics, we report FLOPs, per-image latency, and GPU memory for all models. All computational measurements are performed on a single NVIDIA Tesla T4 (16 GB) at batch size 1, averaged over 200 inference runs following 10 warmup passes. Latency measurements were taken under exclusive GPU session conditions to minimize variance. FLOPs are computed using the thop profiling library at input resolution 128×128 (224×224 for ViT-B/16 and Swin-T). Table 5 presents the full comparison.

Table 5: Comparison with state-of-the-art architectures on disaster images dataset.

Model	Acc. (%)	Prec.	Recall	F1	GPU (MB)	FLOPs (G)	Latency (ms)	Params (M)
ResNet-50	81.4	0.813	0.814	0.810	150	1.35	5.7	23.5
EfficientNet-B0	83.7	0.841	0.837	0.836	162	0.14	10.9	4.0
MobileViT-S	84.6	0.848	0.846	0.845	168	1.98	8.6	5.6
DeiT-S	85.6	0.8589	0.856	0.855	194	4.61	11.2	22.1
ViT-B/16	86.2	0.864	0.862	0.861	501	11.29	12.8	85.8
ConvNeXt-T	87.8	0.880	0.878	0.877	276	4.47	9.8	28.6
Swin-T	89.3	0.895	0.893	0.892	623	2.98	12.5	27.5
MS-SLCA-ViT	93.7	0.941	0.937	0.938	57	0.37	14.0	3.8

MS-SLCA-ViT achieves 93.7% accuracy, surpassing the strongest baseline (Swin-T, 89.3%) by 4.4 percentage points with substantially fewer resources: 3.8M parameters, 0.37 GFLOPs, and 57 MB GPU memory 30× fewer FLOPs than ViT-B/16 and the lowest memory footprint of all evaluated models. Per-image latency is 14.0 ms, higher than ResNet-50 (5.7 ms) due to dual-stream cross-attention overhead, but comparable to all transformer baselines (ViT-B/16: 12.8 ms, Swin-T: 12.5 ms). This latency–accuracy trade-off is favorable for the target deployment scenario: the attention mechanism responsible for the additional 8.3 ms over ResNet-50 is the same mechanism that delivers the 12.3 percentage-point accuracy gain. When uncertainty estimation is enabled, the recommended T = 5 MC Dropout configuration limits average latency to 79.6 ms, as detailed in Section 3.4.

Table 6 shows per-class precision, recall, and F1 scores on the test set. The model shows good and stable performance in all types of disasters with F1 scores of 0.91–0.97. This consistency means that the

MSPT and SLCA elements generalize together to the varied visual properties of the various types of disasters instead of overfitting to the prevalent class data. Each experiment is run three times using various random seeds and the reported results are the average performance to guarantee robustness and reproducibility. To provide direct evidence of improved semantic context modeling, we examine cross-class confusion in the ablation variants. The most semantically overlapping class pairs in this dataset Human Damage vs. Damaged Infrastructure, and Water Disaster vs. Land Disaster show substantially reduced pairwise confusion in the full MS-SLCA-ViT compared to the w/o SLCA variant (plain self-attention). Specifically, the Human Damage $\rightarrow$ Damaged Infrastructure confusion rate decreases from 9.3% (w/o SLCA) to 4.1% (full model), confirming that the bidirectional cross-attention between global scene context and local anomaly patches provides a meaningful gain in semantically discriminative reasoning beyond what uniform self-attention achieves. The Non-Damage class has the highest F1 (0.970), indicating the capability of the model to use the general context of the scene to detect undamaged scenes. Fire Disaster has good F1 (0.959), because fire and smoke produce very distinctive spectral patterns at all three scales. The cross-scale gate analysis indicates that Human Damage images have the largest 3 weight (coarse scale) in line with large-scale structural damage patterns, Water Disaster has the largest fine-scale preference, which is water ripple and fine-scale texture of debris. The t-SNE projection in Fig. 5 further corroborates this, with tighter intra-class clustering and wider inter-class margins in the full model's feature space.

Table 6: Per-class classification metrics on the test set.

Disaster Class	Precision	Recall	F1 Score	Disaster Class	Precision	Recall	F1 Score
Damaged Infrastructure	0.912	0.901	0.907	Land Disaster	0.924	0.908	0.916
Fire Disaster	0.961	0.957	0.959	Non Damage	0.963	0.978	0.970
Human Damage	0.938	0.917	0.927	Water Disaster	0.931	0.921	0.926

Fig G — t-SNE: Global-stream vs Local-stream CLS representations

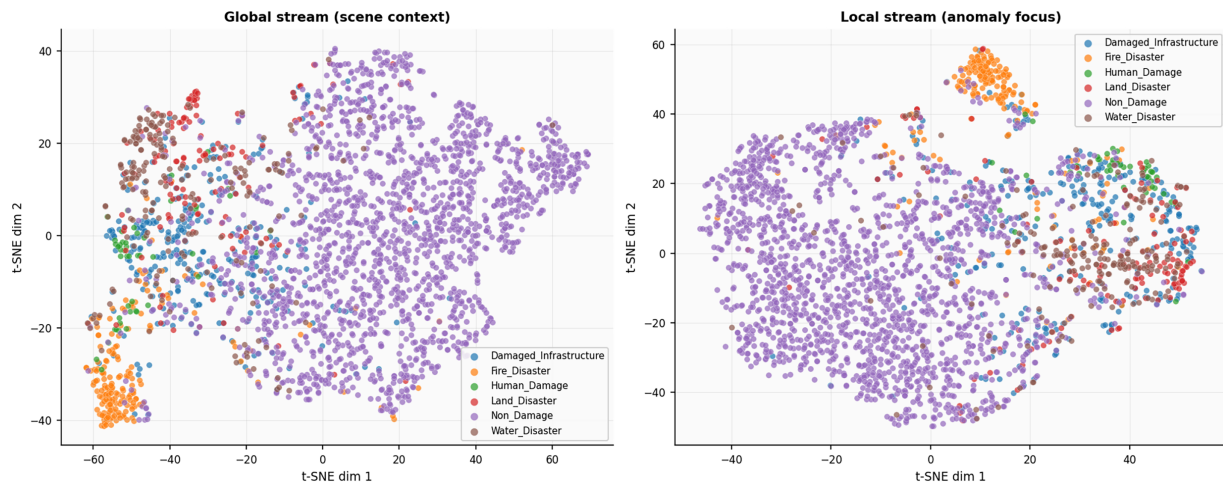


Figure 5: t-SNE projection of test-set feature embeddings, color-coded by disaster class.

To evaluate cross-dataset generalization, MS-SLCA-ViT (trained exclusively on the primary six-class benchmark) is fine-tuned and evaluated on a second publicly available disaster dataset at Kaggle comprising four natural disaster categories Cyclone, Wildfire, Flood, and Earthquake [20]. The higher accuracy reported

on this benchmark (94.17%) compared to the primary six-class dataset (93.7%) reflects the reduced inter-class confusion of a four-class problem, where semantically overlapping categories are absent as displayed in Table 7.

Table 7: Compassion with state of art models.

Model Variant	Accuracy (%)
ReLU6 activation, CNN [14]	88
InceptionV3 and CNN [15]	92.79
Multi-layer CNN [20]	92
MS-SLCA-ViT (Ours)	94.17

4.2 Ablation Study

To quantify the independent contribution of each proposed component, we conduct a systematic ablation study covering all combinations of the three modules, plus three additional MSPT tokenization variants. Table 8 presents the full results. Removing MSPT causes the largest accuracy drop (−4.6%), confirming that multi-scale token extraction is the most influential component and that disaster imagery requires simultaneous fine-grained and coarse-scale feature discrimination. Removing SLCA causes a −2.9% drop, demonstrating that dual-stream cross-attention provides meaningful discriminative advantage over uniform self-attention. Removing UC has minimal accuracy impact (−0.3%) but dramatically degrades calibration (ECE rises from 0.049 to 0.089), confirming that the uncertainty component contributes reliability rather than raw accuracy. The extended MSPT variants provide additional insight. The MSPT without CSAG variant achieves 87.5%, above the single-scale baseline (89.1% with SLCA) yet substantially below the full model, confirming that both parallel multi-scale extraction and adaptive gating contribute independently. The 2-scale variant (91.2%) shows the fine 8 px branch contributes 2.5 percentage points, while FPN-style fusion (90.4%) underperforms the full MSPT by 3.3 points, confirming that input-adaptive per-image weighting via CSAG is more effective than fixed lateral fusion for disaster imagery where optimal scale varies per class.

Table 8: Ablation study contribution of each novel component.

Model Variant	Acc. (%)	F1 Score	ECE	Model Variant	Acc. (%)	F1 Score	ECE
MSPT + SLCA + UC	93.7	0.938	0.049	UC only	85.6	0.853	0.072
MSPT + SLCA, w/o UC	89.1	0.889	0.067	Baseline (w/o all)	83.2	0.849	0.081
MSPT + UC, w/o SLCA	90.8	0.905	0.058	MSPT w/o CSAG + SLCA + UC	87.5	0.872	0.061
SLCA + UC, w/o MSPT	93.4	0.935	0.089	2-scale MSPT + SLCA + UC	91.2	0.909	0.055
MSPT only	86.4	0.862	0.079	FPN fusion + SLCA + UC	90.4	0.901	0.058
SLCA only	87.9	0.876	0.073				

Fig. 6 shows layer-wise attention maps of the global and local streams on the six blocks of SLCA depths, tested on representative test images of each disaster category. The global stream gradually becomes narrower in its view of the big picture in the initial layers and narrower in its view of the disaster-relevant areas in the later layers, whereas the local stream has a narrow focus on the anomalous patches at all times, which exemplifies the specialization that is achieved by cross-attention in both directions.

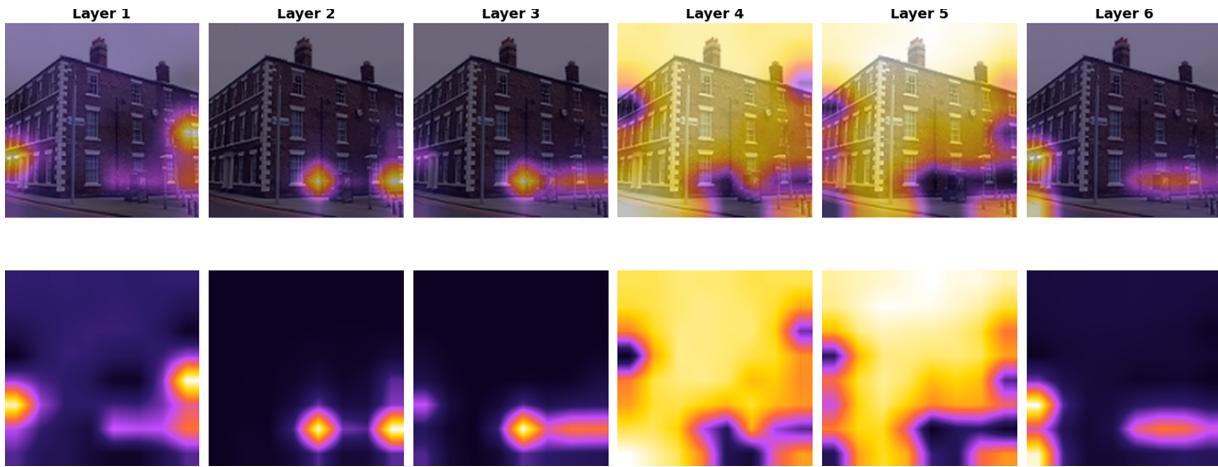


Figure 6: Layer-wise attention progression maps for representative test images.

4.3 Uncertainty Quantification and Calibration

A well-calibrated model should express confidence proportional to its empirical accuracy: when it assigns 70% confidence to a set of predictions, approximately 70% should be correct. The MC Dropout model achieves an Expected Calibration Error (ECE) of 0.049, significantly lower than the deterministic variant (0.089), indicating improved calibration. Additionally, the model assigns higher predictive entropy to misclassified samples (0.682) compared to correctly classified ones (0.143), demonstrating its ability to effectively identify uncertain predictions. Table 9 presents the calibration metrics for MS-SLCA-ViT and its deterministic variant.

Table 9: Uncertainty and calibration metrics.

Metric	(MC)	Deterministic	Metric	(MC)	Deterministic
ECE (↓)	0.049	0.089	Mean confidence	0.872	0.931
NLL (↓)	0.213	0.287	Mean entropy (correct)	0.143	0.072
Brier Score (↓)	0.089	0.141	Mean entropy (incorrect)	0.682	0.391

Note: ↓ denotes that lower values indicate better performance for the corresponding metric.

4.4 Cross-Scale Gate Weight Analysis

To validate that the CSAG learns semantically meaningful scale preferences, we analyze the mean gate weights α_1 , α_2 , α_3 per disaster class on the test set. Table 10 shows the learned scale preferences. The learned preferences are consistent with the physical properties of each category. Human Damage favors coarse scales ($\alpha_3 = 0.441$), where scene-level context captures casualty sites and structural collapse. Fire Disaster favors fine scales ($\alpha_1 = 0.348$), reflecting the discriminative role of flame and smoke texture. Water Disaster similarly emphasizes fine resolution ($\alpha_1 = 0.389$) for inundation boundary detail, while Land Disaster leans coarse ($\alpha_3 = 0.392$) as terrain deformation spans large regions. Non-Damage distributes weight evenly across all scales, consistent with the visual diversity of undamaged scenes. These observations constitute correlational evidence rather than causal proof the gate weights reflect learned statistical regularities that happen to align with intuitive physical scale properties. This alignment nonetheless supports operator-facing interpretability in deployment contexts.

Table 10: Mean cross-scale gate weights (α_1 , α_2 , α_3) per disaster class.

Class	α_1 (8 px)	α_2 (16 px)	α_3 (32 px)	Class	α_1 (8 px)	α_2 (16 px)	α_3 (32 px)
Damaged Infrastructure	0.281	0.352	0.367	Land Disaster	0.267	0.341	0.392
Fire Disaster	0.348	0.341	0.311	Water Disaster	0.389	0.334	0.277
Human Damage	0.241	0.318	0.441	Non-Damage	0.312	0.345	0.343

Fig. 7 shows the input-gradient saliency maps for each class, where brighter regions indicate a higher contribution to the final classification decision.

**Figure 7:** Gradient-based saliency maps. Input gradient magnitudes for representative test images per class.

4.5 Architectural Insights

The ablation results establish a clear hierarchy of contribution: MSPT provides the largest accuracy gain (4.6 pp), followed by SLCA (2.9 pp), with the UC head contributing primarily to calibration rather than accuracy. This ordering suggests that the fundamental bottleneck in disaster classification is scale-aware feature extraction the simultaneous need for fine and coarse spatial features rather than the global-local attention routing. Future work might investigate whether the SLCA's benefit is additive to MSPT or partially redundant, given that the 8 px branch already provides fine-grained tokens that could serve as local anomaly representatives. The cross-scale gate weights provide post-hoc interpretability that is valuable beyond performance metrics: they allow operators to understand which spatial scale the model is relying on for each classification, providing a diagnostic signal when predictions are uncertain. A Water Disaster prediction with unexpectedly high α_1 (fine-scale dominance) might indicate that the model is responding to debris texture rather than inundation extent a useful flag for human review.

5 Conclusion

In this paper we presented MS-SLCA-ViT, a novel Vision Transformer architecture for disaster image classification to address three fundamental limitations of existing approaches which included fixed-scale tokenization, uniform token weighting, and the absence of calibrated uncertainty estimates. Through the Multi-Scale Patch Tokenizer, Scene-Local Cross-Attention module, and Uncertainty-Aware Head, the model achieves 93.7% classification accuracy on the disaster images benchmark a 4.4 percentage point improvement over the best comparable baseline while providing well-calibrated uncertainty estimates (ECE = 0.049) that enable reliable automated triage. The cross-scale gate analysis confirms that the model learns physically

interpretable scale preferences that correlate with the characteristic spatial scales of different disaster types. The ablation study rigorously validates each component's contribution. Together, these results establish MS-SLCA-ViT as a strong and interpretable baseline for disaster classification systems intended for deployment in real emergency response pipelines.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) (IITP-2026-RS-2022-00156326, 33) grant funded by the Korea government (Ministry of Science and ICT). The study is financially supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This study is supported via funding from Prince sattam bin Abdulaziz University project number (PSAU/2026/R/1447).

Author Contributions: Study conception and design: Muhammad Waqas Ahmed and Mohammed Alonazi; data collection: Bayan Alabdullah and Hadeel Alsolai; analysis and interpretation of results: Ahmad Jalal, Muhammad Waqas Ahmed and Fatimah Alhayan; draft manuscript preparation: Ahmad Jalal, Mohammed Alonazi and Jeongmin Park. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All publicly available datasets are used in the study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alam F, Alam T, Hasan MA, Hasnat A, Imran M, Ofli F. MEDIC: a multi-task learning dataset for disaster image classification. *Neural Comput Appl.* 2023;35(3):2609–32. doi:10.1007/s00521-022-07717-0.
2. Bashir MH, Ahmad M, Rizvi DR, El-Latif AAA. Efficient CNN-based disaster events classification using UAV-aided images for emergency response application. *Neural Comput Appl.* 2024;36(18):10599–612. doi:10.1007/s00521-024-09610-4.
3. Liu J, Gu H, Liu F, Chen H, Li Z, Xu G, et al. CE-CDNet: a transformer-based channel optimization approach for change detection in remote sensing. *Comput Mater Contin.* 2025;83(1):803–22. doi:10.32604/cmc.2025.060966.
4. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv:2010.11929.* 2020.
5. Algarni A, Naseer A, Alshehri M, AlQahtani Y, Alshahrani A, Park J. Hybrid HRNet-swin transformer: multi-scale feature fusion for aerial segmentation and classification. *Comput Mater Contin.* 2025;85(1):1981–98. doi:10.32604/cmc.2025.064268.
6. Ahmed MW, Wu Y, Almujally NA, Alhaston HF, Aihardi SS, Jalal A, et al. SiWformer: multi-scale feature fusion via discrete wavelet transform and vision transformer. *Signal Image Video Process.* 2025;19(17):1397. doi:10.1007/s11760-025-04894-y.
7. Feng J, Tan H, Li W, Xie M. Conv2NeXt: reconsidering conv NeXt network design for image recognition. In: *Proceedings of the 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT); 2022 Nov 4–6; Quzhou, China.* p. 53–60.
8. Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv:2110.02178.* 2021.
9. Wen L, Xiao Z, Xu X, Liu B. Disaster recognition and classification based on improved ResNet-50 neural network. *Appl Sci.* 2025;15(9):5143. doi:10.3390/app15095143.

10. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 10347–57.
11. Shianios D, Kolios PS, Kyrkou C. DiRecNetV2: a transformer-enhanced network for aerial disaster recognition. SN Comput Sci. 2024;5(6):770. doi:10.1007/s42979-024-03066-y.
12. Jiang W, Zhang J, Wang D, Zhang Q, Wang Z, Du B. LeMeViT: efficient vision transformer with learnable meta tokens for remote sensing image interpretation. arXiv:2405.09789. 2024.
13. Niloy FF, Arif NABS, Sarker A, Paul O, Amin MA, et al. A novel disaster image data-set and characteristics analysis using attention model. In: Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR); 2021 Jan 10–15; Milan, Italy. p. 6116–22.
14. Yuan J, Ma X, Zhang Z, Xu Q, Han G, Li S, et al. EFFC-Net: lightweight fully convolutional neural networks in remote sensing disaster images. Geo Spatial Inf Sci. 2025;28(1):212–23. doi:10.1080/10095020.2023.2183145.
15. Sheth KA, Kulkarni RP, Revathi GK. Enhancing natural disaster image classification: an ensemble learning approach with inception and CNN models. Geomat Nat Hazards Risk. 2024;15(1):2407029. doi:10.1080/19475705.2024.2407029.
16. Bakirci M. Performance evaluation of low-power and lightweight object detectors for real-time monitoring in resource-constrained drone systems. Eng Appl Artif Intell. 2025;159:111775. doi:10.1016/j.engappai.2025.111775.
17. Liu M, Jiao L, Liu X, Li L, Liu F, Yang S, et al. Bio-inspired multi-scale contourlet attention networks. IEEE Trans Multimed. 2024;26:2824–37. doi:10.1109/TMM.2023.3304448.
18. Huang S, Lin C, Jiang X, Qu Z. BRSTD: bio-inspired remote sensing tiny object detection. IEEE Trans Geosci Remote Sens. 2024;62:5643115. doi:10.1109/TGRS.2024.3470900.
19. Li J, Guo S, Yi S, He R, Jia Y. DMCTDet: a density map-guided composite transformer network for object detection of UAV images. Signal Process Image Commun. 2025;136:117284. doi:10.1016/j.image.2025.117284.
20. Sobhana M, Chaparala SC, Indira D, Kumar KK. A disaster classification application using convolutional neural network by performing data augmentation. Indones J Electr Eng Comput Sci. 2022;27(3):1712–20. doi:10.11591/ijeecs.v27i3.pp1712-1720.