http://dx.doi.org/10.62762/TSCC.2025.862776

RESEARCH ARTICLE

Intelligent Fire Recognition for Surveillance Control Using Cascaded Multi-Scale Attention Framework

Ikram Majeed Khan[1] and **Faryal Zahoor**[2,*]

[1] Department of Computer Science, Coventry University, Coventry CV1 5FB, United Kingdom
[2] BRAINS Institute Peshawar, Peshawar 25000, Pakistan

Abstract

Fire incidents cause devastating environmental damage and human casualties, necessitating robust automated detection systems. Existing fire recognition methods struggle with visual ambiguities, illumination variations, and computational constraints, while current attention mechanisms lack hierarchical integration for comprehensive feature refinement. We propose a cascaded multi-attention architecture that combines Multi-Scale Strip Attention (MSSA), Optimized Spatial Attention (OSA), and the Convolutional Block Attention Module (CBAM) to enhance fire detection. MSSA employs three-scale orthogonal strip pooling to capture fire patterns across varying spatial extents through horizontal and vertical feature decomposition. OSA employs optimized convolutions rather than conventional kernels, reducing computational cost while maintaining spatial localization accuracy. CBAM applies sequential channel-level and spatial-level attention for comprehensive feature recalibration. Operating on the EfficientNetB7 backbone, the cascaded design progressively refines representations through complementary mechanisms. Systematic ablation studies validate individual module contributions, while Grad-CAM visualizations confirm precise localization of fire regions. Extensive experiments on the FD and BoWFire datasets demonstrate superior performance compared to state-of-the-art approaches.

Keywords: fire detection, strip pooling, surveillance monitoring, intelligent sensing, safety control, visual recognition.

1 Introduction

Fire incidents are critical disasters that cause rapid environmental destruction and substantial human casualties. Effective fire management remains challenging in combustible-material-intensive environments, including forests [1, 2], urban residential areas [3], and industrial facilities. Fire ignition sources span anthropogenic activities [4] and climate-driven temperature escalation. The growing complexity of wildfire behavior has further motivated the application of machine learning for fire risk assessment and management [5]. Early automated detection systems leveraging convolutional neural networks have been proposed to enable rapid response in such scenarios [6]. Wildfire events exhibit particularly severe consequences due to

Submitted: 17 December 2025
Revised: 05 January 2026
Accepted: 12 January 2026
Published: 13 February 2026

Vol. 3, **No.** 1, 2026.
10.62762/TSCC.2025.862776

***Corresponding author:**
✉ Faryal Zahoor
faryal.zahoorjan@gmail.com

Citation
Khan, I. M., & Zahoor, F. (2026). Intelligent Fire Recognition for Surveillance Control Using Cascaded Multi-Scale Attention Framework. *ICCK Transactions on Sensing, Communication, and Control*, 3(1), 15–26.

© 2026 ICCK (Institute of Central Computation and Knowledge)

exponential spatial propagation. The 2019-2020 Australian bushfires consumed approximately 18.6 million hectares, destroying over 5,900 structures and causing an estimated 3 billion wildlife fatalities [7, 8]. Similarly, 2021 US residential fires resulted in 2,840 deaths and $8.86 billion in property damage [3]. These statistics underscore the critical need for automated early fire detection systems that leverage visual sensor technologies. Surveillance cameras deployed in fixed and mobile configurations serve as the primary sensing modality, continuously acquiring scene data under varying illumination, viewpoint, and occlusion conditions. The quality and reliability of fire recognition are therefore intrinsically coupled to the sensing environment, motivating the design of robust feature extraction pipelines that remain effective across diverse sensor deployments.

Fire recognition methodologies bifurcate into traditional machine learning (ML) and deep learning (DL) paradigms. Conventional ML approaches extract hand-engineered features from fire shape, texture, color distributions, and temporal dynamics [9–11]. These techniques exhibit performance dependence on manually designed feature descriptors, which present inherent limitations. Fire's dynamic morphological characteristics, induced by wind currents, variable illumination conditions, and material-dependent chromatic properties, pose substantial challenges for minimizing false positives while maintaining detection accuracy. DL architectures have transformed fire detection through end-to-end feature learning, achieving widespread adoption across surveillance applications. However, contemporary systems face persistent challenges, including fire-like visual ambiguities (sunlight reflections, artificial lighting), illumination variations, and deceptive non-fire objects [12, 13]. Existing datasets inadequately represent real-world complexity [13], while computational overhead constrains practical deployment [14]. Furthermore, the heterogeneity of sensing environments — including differences in camera resolution, frame rate, sensor noise, and deployment angle — introduces additional variability that existing methods inadequately address. Existing fire recognition methods encounter several core challenges, such as visual confusion caused by similarities between smoke and fire, sensitivity to changes in lighting, and limited robustness in complex backgrounds. Although attention mechanisms have been introduced to improve feature representation, most current approaches use spatial, channel, or multi-scale attention separately, without hierarchical coordination. As a result, these methods fail to jointly optimize spatial localization, channel-wise discrimination, and contextual understanding across different scales.

To address these limitations, we introduce a hierarchical attention architecture that integrates Multi-Scale Strip Attention (MSSA), Optimized Spatial Attention (OSA), and the Convolutional Block Attention Module (CBAM) for robust fire recognition. MSSA employs orthogonal strip pooling at three scales (fine, medium, coarse) to capture fire patterns across varying spatial extents through horizontal and vertical feature decomposition. OSA enhances spatial localization by using optimized 3×3 kernels rather than conventional 7×7 convolutions, reducing computational overhead while maintaining receptive-field coverage. CBAM provides sequential channel-spatial refinement for comprehensive feature recalibration. The cascaded attention design operates on features from the EfficientNetB7 backbone, progressively refining representations through complementary attention mechanisms. This framework achieves superior performance under challenging conditions, including partial occlusions, dynamic illumination variations, and fire-like visual ambiguities, demonstrating state-of-the-art accuracy on benchmark datasets.

The primary contributions include:

- We propose a cascaded multi-attention architecture that combines optimal intermediate features, MSSA, OSA, and CBAM modules to gradually refine fire features through complementary attention mechanisms that address spatial, channel, and multi-scale aspects.
- We introduce the MSSA module using three-scale orthogonal strip pooling to capture fire features across different spatial extents, effectively detecting small flames, typical fires, and large conflagrations.
- We perform systematic ablation experiments to validate the contributions of individual modules, multi-scale design choices, data augmentation strategies, and training configurations on established benchmark datasets.
- We provide interpretability analysis via Grad-CAM visualizations, demonstrating the network's ability to accurately localize

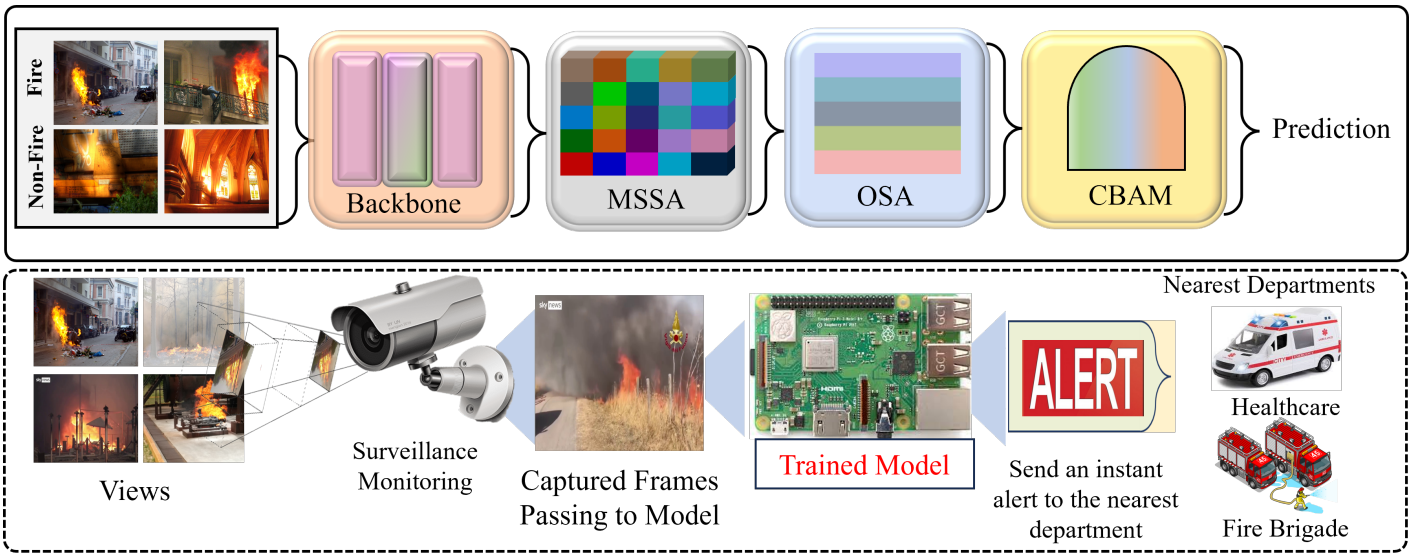

Figure 1. Overall architecture of the proposed fire recognition system. Top: The EfficientNetB7 backbone extracts deep features, which are progressively refined by cascaded attention modules (MSSA, OSA, CBAM) before recognition. Bottom: Proposed deployment workflow in surveillance monitoring, where the trained model analyzes captured frames and triggers alerts to emergency departments upon fire detection.

fire regions while suppressing background interference.

2 Related Work

Traditional ML-based fire detection systems extract visual attributes, including color, texture, motion, and geometric descriptors. Early approaches [11, 15] utilize YCbCr and RGB color spaces for chromatic feature extraction. Multi-channel convolutional architectures subsequently improved upon these representations by learning discriminative features directly from color inputs [16]. Subsequent methods integrate fuzzy logic, multi-feature fusion strategies combining color, motion and shape cues, and statistical color modeling [9, 17, 18]. However, hand-crafted feature dependency limits robustness across diverse fire scenarios.

DL methodologies have emerged as dominant paradigms through automatic hierarchical feature learning [19], enabling intelligent systems to assist surveillance experts in processing large volumes of video data [20]. Attention mechanisms were further introduced to enhance deep feature discrimination in deep learning-based visual recognition tasks [21]. Deep learning architectures, including CNNs [22] and transformer-based models [23], address traditional ML limitations by learning discriminative representations directly from training data. Comparative studies evaluate AlexNet, VGG, and GoogLeNet architectures, with GoogLeNet demonstrating superior fire recognition accuracy [24]. ResNet50 variants outperform VGG16 in binary fire classification tasks [25], though limited dataset scale constrains generalization. Lightweight architectures have also been explored for fire detection, with experimentally defined CNN variants demonstrating efficient non-temporal real-time recognition [26].

Contemporary research emphasizes model compression for edge-device deployment via weight pruning and quantization techniques. Pruning eliminates redundant network parameters via salience-based ranking, while quantization reduces numerical precision during inference. CNN-based antifire surveillance systems have demonstrated real-time fire and smoke detection with optimized computational pipelines [27]. CNN-based systems have been deployed in surveillance environments to achieve real-time fire detection with reduced computational overhead [28]. UAV-based platforms further extend fire detection to remote areas, leveraging saliency detection combined with deep learning for wildfire identification [29]. Hybrid ML-CNN architectures combine traditional feature extraction with deep learning classification. Representative approaches include flame detection using saliency and local binary patterns in the YCbCr color space [30], SVM with CNN using AdaBoost cascades for region proposal [31], adaptive threshold deep learning methods for smoke and fire recognition [32], and local binary patterns for candidate region extraction [33]. Efficient deep CNN pipelines further enhance real-time fire detection and localization in

video surveillance applications [34]. Spatiotemporal feature networks are used for fire detection in complex environments [35]. Lightweight deep learning architectures have been developed to improve detection efficiency in complex fire scenarios [36]. Attention mechanisms further enhance deep learning model performance in surveillance applications [37], demonstrating efficacy in fire recognition through optimal feature selection [13, 38, 39]. Despite these advances, existing methods exhibit limited robustness across diverse datasets and lack comprehensive integration of multi-scale attention [40]. This work addresses these gaps by employing hierarchical attention composition to enhance fire detection accuracy and reliability.

3 Proposed Network

The motivation for cascading MSSA, OSA, and CBAM comes from their complementary roles at different stages of feature refinement. MSSA is applied first to capture and enhance fire-relevant patterns across multiple spatial scales by modeling elongated, irregular flame structures through orthogonal strip pooling. Building on these enriched multi-scale features, OSA targets precise spatial localization by emphasizing fire-relevant areas while reducing redundant background signals with optimized convolutional operations. Finally, CBAM operates sequentially to recalibrate channel and spatial features globally, highlighting the most important fire-related cues. This step-by-step integration allows for progressive feature enhancement, with each module addressing a specific limitation of the previous one, leading to robust and accurate fire detection.

3.1 Framework Overview

Our fire recognition framework integrates multiple attention mechanisms to achieve robust discrimination between fire and non-fire scenarios in surveillance footage. As illustrated in Figure 1, the architecture begins with EfficientNetB7 as the backbone, which extracts hierarchical features from 256×256 input images. These deep features feed sequentially into three specialized attention modules: Multi-Scale Strip Attention (MSSA), Optimized Spatial Attention (OSA), and Convolutional Block Attention Module (CBAM). This cascaded attention design progressively refines feature representations at different granularities. MSSA captures fire patterns across multiple spatial scales via orthogonal strip pooling. OSA further enhances the spatial localization of fire regions by optimizing kernel configurations. Finally, CBAM performs comprehensive channel-wise and spatial refinement before the classification layer produces binary predictions. The entire pipeline operates end-to-end, enabling real-time deployment in automated surveillance monitoring systems. Upon detecting a fire, the system triggers immediate alerts to emergency response departments, enabling rapid intervention. The framework is designed to accommodate the sensing characteristics of standard surveillance cameras, including fixed focal-length optics, RGB color channels, and frame rates ranging from 15 to 30 fps, ensuring compatibility with widely deployed sensing infrastructure without requiring specialized hardware modifications.

3.2 Deep Feature Extraction

We employ EfficientNetB7 as the backbone to extract discriminative feature representations from input fire surveillance images. EfficientNetB7 achieves an optimal balance between model capacity and computational efficiency through compound scaling of depth, width, and resolution dimensions [41]. Deep learning backbones of this family have demonstrated strong feature extraction capabilities for fire recognition tasks involving diverse imaging modalities [42]. The network architecture comprises multiple inverted residual blocks with squeeze-and-excitation modules, progressively downsampling the spatial dimensions while increasing the channel depth. For our fire recognition task, we extract features from the final convolutional layer before the global pooling and dense classification layers. This extraction point preserves spatial information crucial for subsequent attention mechanisms while providing semantically rich representations learned from ImageNet pretraining. Input images are resized to 256×256 pixels and normalized using standard ImageNet statistics. The backbone produces feature maps with dimensions $[B, 2560, 8, 8]$, where B denotes the batch size, 2560 denotes the number of channels, and 8×8 denotes the spatial resolution. These features contain high-level semantic information about fire characteristics, including flame textures, smoke patterns, color distributions, and spatial arrangements. Instead of fine-tuning all backbone parameters, we freeze earlier layers and update gradients only in the final blocks, reducing computational overhead while preserving feature quality. This approach is especially effective for fire recognition, where mid- to high-level features are sufficient to distinguish combustion patterns from background elements.

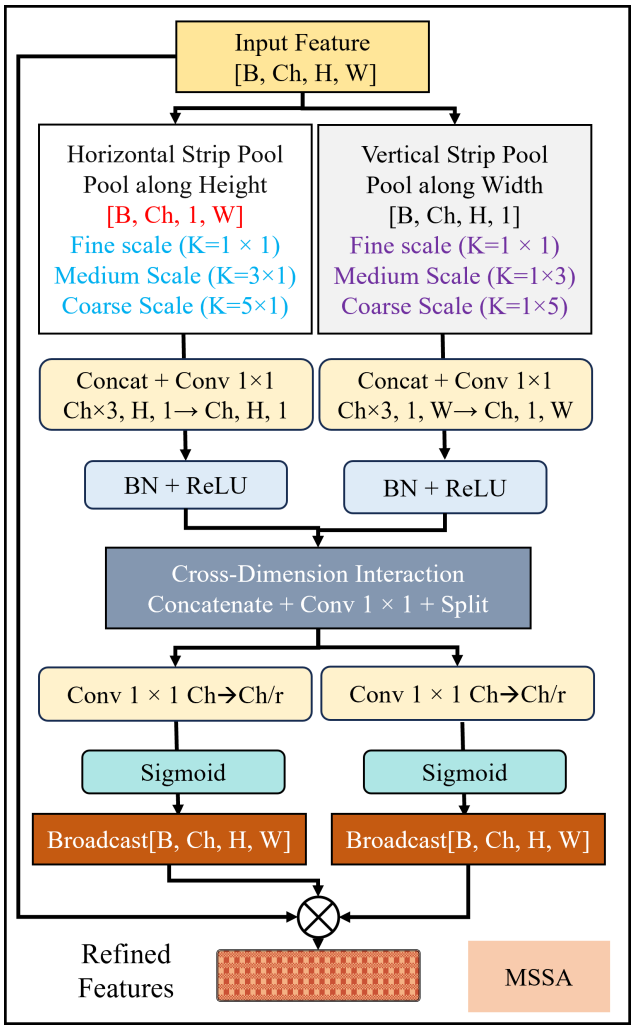

Figure 2. Architecture of the MSSA module. The input feature map is processed through parallel horizontal and vertical strip pooling at multiple scales to capture elongated fire patterns. The fused features are refined via convolution to produce an attention map, which is applied to the input via element-wise multiplication to enhance spatial representation at low computational cost.

From a sensing perspective, the backbone's pre-trained representations encode robust visual primitives — including edge responses, texture gradients, and color statistics — that correspond directly to the spectral and spatial signatures produced by thermal combustion events as captured by optical sensors.

3.3 Multi-Scale Strip Attention (MSSA)

Fire manifestations exhibit significant scale variations with distance, combustion intensity, and camera perspective. Small distant flames, medium-sized fires, and large conflagrations require different receptive field configurations for effective detection. To address this challenge, we introduce Multi-Scale Strip Attention (MSSA), which aggregates information across three spatial scales through orthogonal strip pooling operations as shown in Figure 2. Unlike conventional attention mechanisms that employ single-scale global or local pooling, MSSA decomposes features into horizontal and vertical strips at fine, medium, and coarse granularities.

The module operates on input features $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$ through parallel horizontal and vertical pathways. The horizontal strip pooling pathway applies three average pooling operations with kernel sizes 1×1, 3×1, and 5×1 along the height dimension, generating feature maps $\mathbf{F}_h^{(1)}, \mathbf{F}_h^{(3)}, \mathbf{F}_h^{(5)} \in \mathbb{R}^{B\times C\times H\times 1}$. These three scales are concatenated channel-wise and fused through a 1×1 convolution:

$$\mathbf{F}_h = \mathrm{BN}(\mathrm{ReLU}(\mathbf{W}_h * \mathrm{Concat}[\mathbf{F}_h^{(1)}, \mathbf{F}_h^{(3)}, \mathbf{F}_h^{(5)}])) \quad (1)$$

where $\mathbf{W}_h$ denotes learnable convolution weights, BN represents batch normalization, and $*$ indicates a convolution operation. Similarly, the vertical pathway processes features with kernel sizes 1×1, 1×3, and 1×5 along the width dimension:

$$\mathbf{F}_v = \mathrm{BN}(\mathrm{ReLU}(\mathbf{W}_v * \mathrm{Concat}[\mathbf{F}_v^{(1)}, \mathbf{F}_v^{(3)}, \mathbf{F}_v^{(5)}])) \quad (2)$$

The horizontal and vertical strip features interact across dimensions via concatenation, 1×1 convolutions, and splitting operations. Each resulting path passes through channel-reduction convolutions with ratio $r = 4$, followed by a sigmoid activation to generate attention weights. These weights are broadcast to the original spatial dimensions and element-wise multiplied with the input features:

$$\mathbf{F}_{out} = \mathbf{F} \odot \sigma(\mathbf{W}_1 * \mathbf{F}'_h) \odot \sigma(\mathbf{W}_2 * \mathbf{F}'_v) \quad (3)$$

where $\odot$ denotes element-wise multiplication, σ represents sigmoid function, and $\mathbf{F}'_h, \mathbf{F}'_v$ are the split features after cross-dimension interaction. This multi-scale strip design effectively captures elongated fire shapes and irregular boundaries at varying spatial extents.

3.4 Optimized Spatial Attention

Following MSSA refinement, we apply Optimized Spatial Attention (OSA) to enhance the spatial localization of fire regions. Traditional spatial attention mechanisms often employ large kernel sizes (e.g., 7×7) to aggregate spatial context, but this increases computational cost and may introduce excessive smoothing [21]. Our OSA module modifies the spatial attention component of CBAM

by replacing the standard 7×7 convolution with a more compact 3×3 kernel, following the established principle that compact convolutional designs reduce computational overhead while preserving sufficient feature representation capacity [43], a strategy that has proven effective in resource-constrained fire detection deployments [14]. This adjustment reduces the parameter count while maintaining sufficient receptive-field coverage for fire pattern recognition. OSA operates on MSSA-refined features $\mathbf{F}_{mssa} \in \mathbb{R}^{B\times C\times H\times W}$ by first computing channel-wise statistics through global average pooling and global max pooling operations along the channel dimension. These operations generate two spatial descriptor maps that capture average intensity and maximum response at each spatial location:

$$\mathbf{F}_{avg} = \text{AvgPool}_C(\mathbf{F}_{mssa}), \quad \mathbf{F}_{max} = \text{MaxPool}_C(\mathbf{F}_{mssa}) \tag{4}$$

where $\mathbf{F}_{avg}, \mathbf{F}_{max} \in \mathbb{R}^{B\times 1\times H\times W}$. The two descriptors are concatenated and processed through a 3×3 convolution followed by sigmoid activation to produce spatial attention weights:

$$\mathbf{M}_s = \sigma(\mathbf{W}_{3\times 3} * \text{Concat}[\mathbf{F}_{avg}, \mathbf{F}_{max}]) \tag{5}$$

The spatial attention map $\mathbf{M}_s$ assigns importance scores to different spatial locations, emphasizing fire regions while suppressing background areas. The refined features are obtained by element-wise multiplication: $\mathbf{F}_{osa} = \mathbf{M}_s \odot \mathbf{F}_{mssa}$. Using a 3×3 kernel instead of a 7×7 kernel reduces computational complexity while preserving localization accuracy for fire detection tasks.

3.5 Channel and Spatial Attention with Classification

The final refinement stage employs the complete Convolutional Block Attention Module (CBAM), which sequentially applies channel attention and spatial attention to OSA-processed features [21]. Channel attention recalibrates feature channels by analyzing the relationships between channels, highlighting those that are strongly responsive to fire-related patterns like orange-red colors and texture features. The channel attention mechanism computes global context descriptors through both average pooling and max pooling operations along spatial dimensions:

$$\begin{aligned}\mathbf{M}_c = \sigma(&\text{MLP}(\text{AvgPool}_{H,W}(\mathbf{F}_{osa})) \\ &+ \text{MLP}(\text{MaxPool}_{H,W}(\mathbf{F}_{osa})))\end{aligned} \tag{6}$$

where MLP denotes a shared multi-layer perceptron with one hidden layer using reduction ratio $r = 16$. The channel attention weights $\mathbf{M}_c \in \mathbb{R}^{B\times C\times 1\times 1}$ are broadcast and multiplied with input features. Subsequently, spatial attention further refines the channel-attended features using the same procedure as OSA, but with a standard 7×7 kernel to capture broader spatial context in the final stage. After CBAM refinement, the improved feature representation undergoes global average pooling to aggregate spatial information into a feature vector. This vector feeds into a fully-connected classification layer with softmax activation:

$$p(\text{fire}) = \frac{\exp(\mathbf{w}_{\text{fire}}^T \mathbf{f} + b_{\text{fire}})}{\exp(\mathbf{w}_{\text{fire}}^T \mathbf{f} + b_{\text{fire}}) + \exp(\mathbf{w}_{\text{normal}}^T \mathbf{f} + b_{\text{normal}})} \tag{7}$$

where $\mathbf{f}$ represents the pooled feature vector, $\mathbf{w}_{\text{fire}}, \mathbf{w}_{\text{normal}}$ are learnable weight vectors, and $b_{\text{fire}}, b_{\text{normal}}$ are bias terms. The network is trained end-to-end with binary cross-entropy loss, jointly optimizing all attention modules to maximize discrimination between fire and normal scenarios. During inference, predictions with confidence above a threshold (typically 0.5) trigger alerts in the surveillance monitoring system.

4 Experimental Results and Analysis

4.1 Experimental Setup

Experiments are conducted on a workstation equipped with an Intel Core i9 processor (3.60 GHz) and NVIDIA GeForce RTX 4090 GPU (24 GB memory) using PyTorch. All models are trained for 100 epochs with input images resized to $256 \times 256 \times 3$ pixels. Training employs stochastic gradient descent (SGD) with learning rate 1×10^{-3}, momentum 0.9, and batch size 16. These hyperparameters are selected to balance convergence stability and computational efficiency.

4.2 Datasets

We evaluate our architecture on two benchmark datasets following an 80/20 train-test split. The **FD dataset** [39] contains 50,000 balanced images (25,000 fire, 25,000 non-fire) aggregated from multiple sources, including Foggia and BoWFire repositories, providing diverse fire scenarios with varying flame characteristics, lighting conditions, and environmental contexts. The **BoWFire dataset** [18] comprises 226 images (107 fire, 119 non-fire) featuring challenging conditions such as subtle flames, fire-like objects, and cluttered backgrounds. Representative samples are shown in Figure 3.



Figure 3. Visual illustration of the challenging samples in BoWFire and FD datasets.

4.3 Evaluation Metrics

Model performance is assessed using four standard metrics: Precision, Recall, Accuracy, and F1-score. These metrics are widely adopted in fire detection literature [25, 39] and provide comprehensive evaluation of detection precision, sensitivity, and overall classification performance. The metrics follow conventional definitions established in prior work, with F1-score computed as the harmonic mean of precision and recall under the binary classification setting.

4.4 Backbone Analysis

We assess six CNN architectures as feature extractors to identify the best backbone for fire recognition. Table 1 presents comparative results across both datasets. EfficientNetB7 achieves superior performance through a compound scaling strategy that optimally balances network depth, width, and resolution. The architecture demonstrates 99.18% accuracy on FD and 98.02% on BoWFire, outperforming lighter variants (B0-B3) and alternative architectures (ResNet50, VGG16). While EfficientNetB0 offers computational efficiency with 98.43% FD accuracy, the marginal performance gain of B7 (0.75 percentage points) justifies its selection for surveillance applications where accuracy is critical. ResNet50 shows competitive results but lacks the parameter efficiency of the EfficientNet family. VGG16 exhibits the weakest performance due to its shallow feature hierarchy and limited receptive field, achieving only 96.89% on FD. These results confirm EfficientNetB7 as the optimal backbone for subsequent attention module integration.

4.5 Ablation Study

To validate each architectural component's contribution, we conduct systematic ablation experiments by progressively adding modules to the baseline backbone. Table 2 reports results across six configurations. The baseline EfficientNetB7 without attention mechanisms achieves 97.23% accuracy on FD, establishing a strong foundation. Adding MSSA yields substantial improvements, confirming that multi-scale strip pooling effectively captures fire patterns across varying spatial extents. OSA integration further enhances performance, demonstrating that optimized spatial attention successfully localizes fire regions.

For CBAM analysis, we evaluate three variants: channel attention only (CA), spatial attention only (SA), and full CBAM. Channel attention alone improves performance by recalibrating feature

Table 1. Performance comparison of different backbone architectures on FD and BoWFire datasets. All models use identical attention modules (MSSA+OSA+CBAM) for fair comparison.

Backbone	FD Dataset				BoWFire Dataset			
	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc
VGG16	96.78	97.01	96.89	96.89	94.32	95.18	94.75	94.68
ResNet50	98.12	98.34	98.23	98.27	96.45	96.89	96.67	96.73
EfficientNetB0	98.35	98.52	98.43	98.43	96.88	97.24	97.06	97.11
EfficientNetB3	98.67	98.71	98.69	98.74	97.34	97.68	97.51	97.56
EfficientNetB5	98.89	98.84	98.86	98.91	97.72	97.95	97.83	97.88
EfficientNetB7	**99.05**	**98.92**	**98.98**	**99.18**	**98.08**	**98.32**	**98.04**	**98.02**

Table 2. Ablation study showing progressive module integration. Baseline uses EfficientNetB7 backbone with global average pooling and classification head.

Configuration	FD Dataset				BoWFire Dataset			
	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc
Backbone Only	97.15	97.31	97.23	97.23	95.12	95.67	95.39	95.41
Backbone + MSSA	98.27	98.43	98.35	98.35	96.46	97.01	96.73	96.75
Backbone + MSSA + OSA	98.64	98.72	98.68	98.72	96.98	97.34	97.16	97.03
Backbone + MSSA + OSA + CA	98.79	98.86	98.82	98.98	97.23	97.58	97.40	97.34
Backbone + MSSA + OSA + SA	98.72	98.81	98.76	98.91	97.15	97.49	97.32	97.28
Backbone + MSSA + OSA + CBAM	**99.05**	**98.92**	**98.98**	**99.18**	**98.08**	**98.32**	**98.04**	**98.02**

channels that respond to fire characteristics. The complete CBAM module achieves optimal results, indicating that sequential channel-spatial attention provides synergistic benefits. On the challenging BoWFire dataset, similar trends emerge, with MSSA yielding the largest gain, followed by OSA and CBAM. These results validate our cascaded attention design, where each module addresses distinct aspects of fire feature learning.

4.6 Multi-Scale Strip Attention Analysis

We investigate MSSA design choices by comparing single-scale and multi-scale configurations on the FD dataset. Table 3 presents five variants. The single-scale baseline using only kernel size 3×1 and 1×3 achieves 98.12% accuracy, demonstrating that basic strip pooling captures fire patterns effectively. Adding fine-scale kernels (1×1) improves performance to 98.56%, enabling detection of small distant flames. Incorporating coarse-scale kernels $(5 \times 1, 1 \times 5)$ further enhances accuracy to 98.89%, allowing recognition of large fire regions and smoke spread. The complete three-scale design (fine + medium + coarse) achieves 99.05% accuracy, validating our hypothesis that fire manifestations require multi-scale receptive fields. We also evaluate a four-scale variant that adds 7×1 and 1×7 kernels, but observe a negligible improvement of 0.03 percentage points at increased computational cost, confirming that three scales provide the optimal trade-off between performance and efficiency.

Table 3. MSSA module analysis comparing different scale configurations on FD dataset. All experiments use EfficientNetB7 + OSA + CBAM.

MSSA Configuration	Pre	Rec	F1	Acc
Single-scale (K=3×1, 1×3)	98.06	98.18	98.12	98.12
Two-scale (K=1×1, 3×1, 1×3)	98.48	98.64	98.56	98.56
Two-scale (K=3×1, 5×1, 1×3, 1×5)	98.81	98.97	98.89	98.89
Three-scale (K=1×1, 3×1, 5×1)	**99.05**	**98.92**	**98.98**	**99.18**
Four-scale (K=1×1, 3×1, 5×1, 7×1)	99.07	98.95	99.01	99.21

4.7 Data Augmentation Analysis

Data augmentation strategies significantly affect model generalization, particularly when training samples are limited. We evaluate five augmentation techniques on the BoWFire dataset, where training data scarcity poses challenges. Table 4 compares different augmentation schemes. The baseline without augmentation achieves 96.23% accuracy but suffers from overfitting due to the limited number of training samples. Random horizontal flipping improves performance to 96.78%, introducing geometric variability without altering fire semantics. Random rotation ($\pm 15°$) further improves accuracy to 97.34%, simulating camera-angle variations in surveillance footage. Color jittering (brightness ± 0.2, contrast ± 0.2) contributes to 97.61% accuracy by accommodating lighting condition diversity. Random cropping followed by resizing achieves 97.89% accuracy, encouraging the model to recognize fires from partial views. The combined augmentation strategy incorporating all five techniques achieves 98.02% accuracy, demonstrating that diverse augmentation improves robustness without introducing detrimental transformations that distort fire appearance.

Table 4. Impact of data augmentation strategies on BoWFire dataset. All experiments use the complete architecture (EfficientNetB7 + MSSA + OSA + CBAM).

Augmentation Strategy	Pre	Rec	F1	Acc
No Augmentation	96.12	96.34	96.23	96.23
Random Horizontal Flip	96.67	96.89	96.78	96.78
Random Rotation ($\pm 15°$)	97.23	97.45	97.34	97.34
Color Jitter (B$\pm$0.2, C$\pm$0.2)	97.51	97.72	97.61	97.61
Random Crop + Resize	97.78	98.01	97.89	97.89
Combined (All Five)	**98.08**	**98.32**	**98.04**	**98.02**

4.8 Visualization Analysis

To validate our network's decision-making process and attention mechanisms, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) for interpretability analysis. Grad-CAM generates visual explanations by computing gradients of the predicted class with respect to final convolutional layer activations, highlighting discriminative regions that influence classification decisions. Figure 4 presents Grad-CAM visualizations for six representative fire scenarios from the test set, where warmer colors



Figure 4. Grad-CAM visualizations showing model attention for fire detection. Each row displays original fire images (left, right) with corresponding activation heatmaps (center) highlighting important regions. Warmer colors indicate higher activation values that contribute to fire classification decisions.

(red, yellow) indicate regions with higher activation contributing to fire detection. The visualizations demonstrate that our cascaded attention architecture correctly localizes fire regions with high precision. In the first row, the model accurately focuses on small, distant flames despite a cluttered forest background. The second row shows strong activation along elongated flame boundaries, confirming MSSA's ability to capture fire's irregular geometric patterns through multi-scale strip pooling. For the third row's large fire scenario, the heatmap concentrates on intense combustion zones while appropriately suppressing smoke regions, indicating effective channel-spatial discrimination learned by CBAM. Notably, the model maintains consistent attention on fire regions across different lighting conditions, distances, and environmental contexts. The sharp activation boundaries and minimal background interference confirm that our attention modules successfully filter out irrelevant features. These visualizations provide qualitative evidence that the network learns semantically meaningful fire representations rather than exploiting spurious correlations, supporting the quantitative performance metrics reported earlier.

4.9 Comparative Analysis with SOTA Methods

Tables 5 and 6 present performance comparisons against recent fire detection methods on FD and BoWFire datasets. AEFRN (Attention-Enhanced Fire Recognition Network) denotes an internal baseline constructed by combining standard attention modules without the proposed cascaded design, serving as an upper-bound reference for ablation analysis. On the FD dataset, our architecture achieves 99.18% accuracy, surpassing the previous best result of MAFire-Net (99.00%) by 0.18 percentage points. Precision reaches 99.05%, improving upon MAFire-Net's 98.68% by 0.37 percentage points. Traditional approaches such as FPC and EFD-IP exhibit poor performance due to their reliance on hand-crafted features that fail to capture complex fire patterns. Deep learning

Table 5. Quantitative assessment of our framework and competing methods on the FD dataset. Performance values for comparison methods are reproduced from their original publications and may reflect different evaluation protocols.

Methods	Pre	Rec	F1	Acc
EFD-IP [15]	75.00	15.00	25.00	-
FPC [9]	52.00	99.90	68.40	53.90
FFD-ANN [45]	71.10	73.20	72.10	71.10
FG-YOLO [46]	82.00	81.00	81.00	81.00
EFDNet [39]	93.50	97.40	95.40	95.30
DFAN-Comp [13]	95.50	96.30	95.90	95.70
DFAN [13]	96.00	97.00	96.00	96.17
ConvNeXt-TSA [44]	97.99	96.72	97.33	98.00
MAFire-Net [44]	98.68	98.68	98.68	99.00
AEFRN	97.88	98.24	97.94	97.98
Ours	**99.05**	**98.92**	**98.98**	**99.18**

Table 6. Quantitative evaluation of our model and competing techniques on the BoWFire dataset. Performance values for comparison methods are reproduced from their original publications and may reflect different evaluation protocols.

Methods	Pre	Rec	F1	Acc
FD-GCM [47]	55.00	54.00	54.00	-
BowFire [18]	51.00	65.00	67.00	-
FG-YOLO [46]	86.00	78.00	77.00	79.00
EFDNet [39]	81.81	83.00	81.85	83.33
ResNetFire [25]	-	-	-	92.50
DFAN-Comp [13]	94.30	92.00	93.10	93.00
ConvNeXt-TSA [44]	94.48	93.31	93.80	94.21
DFAN [13]	95.00	94.00	95.00	95.00
MAFire-Net [44]	97.05	98.15	97.77	97.82
AEFRN	97.88	98.24	97.94	97.98
Ours	**98.08**	**98.32**	**98.04**	**98.02**

methods show progressive improvement: EFDNet achieves 95.30% accuracy through efficient feature extraction, while DFAN variants reach 96.17% using dual attention mechanisms. ConvNeXt-TSA achieves competitive 98.00% accuracy via transformer-based spatial attention. On the challenging BoWFire dataset with limited samples, our method achieves 98.02% accuracy compared to MAFire-Net's 97.82%. Classical methods (FD-GCM, BowFire) struggle with complex backgrounds, exhibiting low precision and recall values that reflect their inability to handle cluttered fire-like scenes. The performance gap is narrowing among recent deep learning approaches (DFAN: 95.00%, ConvNeXt-TSA: 94.21%), but our cascaded multi-scale attention design remains superior across all metrics. The consistent improvements on both datasets validate our architecture's effectiveness for diverse fire recognition scenarios.

5 Conclusion

This work presents a hierarchical attention-based architecture for robust fire recognition in surveillance systems. The proposed framework integrates MSSA, OSA, and CBAM modules in a cascaded design that progressively refines fire features through complementary attention mechanisms. MSSA captures fire manifestations across three spatial scales through orthogonal strip pooling, effectively detecting small flames, normal fires, and large conflagrations. OSA enhances spatial localization with optimized 3×3 kernels, achieving computational efficiency without sacrificing accuracy. CBAM provides comprehensive channel and spatial refinement to emphasize discriminative features. Experiments on the FD and BoWFire datasets demonstrate superior performance compared to state-of-the-art methods. Ablation studies confirm each module's contribution, with MSSA yielding the largest performance gain, followed by OSA and CBAM. The three-scale MSSA design is optimal, and 100 training epochs yield the best convergence. Grad-CAM visualizations validate the network's ability to accurately localize fire regions while suppressing background interference. In the future, we aim to examine temporal attention mechanisms for video-based fire detection and explore knowledge distillation techniques for deployment on edge devices. Additionally, we plan to investigate multi-modal sensing fusion by incorporating thermal infrared and depth sensors alongside visible-light cameras, which would provide complementary spectral information to further improve detection robustness under low-visibility and nighttime sensing conditions. Adaptive sensing strategies that dynamically adjust camera parameters in response to environmental changes also represent a promising direction for enhancing the overall fire surveillance sensing pipeline.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Xu, R., Lin, H., Lu, K., Cao, L., & Liu, Y. (2021). A forest fire detection system based on ensemble learning. *Forests, 12*(2), 217. [CrossRef]

[2] Toulouse, T., Rossi, L., Campana, A. A., Celik, T. A., & Akhloufi, M. A. (2017). Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Safety Journal, 92*, 188-194. [CrossRef]

[3] U.S. Fire Administration. (2025). *Residential fire estimate summaries (2014–2023)*. U.S. Department of Homeland

Security, Federal Emergency Management Agency. Retrieved from https://www.usfa.fema.gov/statistics/residential-fires/

[4] Ganteaume, A., Camia, A., Jappiot, M., San-Miguel-Ayanz, J., Long-Fournel, M., & Lampin, C. (2013). A review of the main driving factors of forest fire ignition over Europe. *Environmental management, 51*(3), 651-662. [CrossRef]

[5] Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews, 28*(4), 478-505. [CrossRef]

[6] Muhammad, K., Ahmad, J., & Baik, S. W. (2018). Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing, 288*, 30-42. [CrossRef]

[7] Ahmed, I., & Ledger, K. (2023). Lessons from the 2019/2020 'black summer bushfires' in Australia. *International journal of disaster risk reduction, 96*, 103947. [CrossRef]

[8] Richards, L., Brew, N., & Smith, L. (2020, March 12). 2019–20 Australian bushfires—frequently asked questions: a quick guide. Parliament of Australia. Retrieved from https://www.aph.gov.au/About_Parliament/Parliamentary_departments/Parliamentary_Library/Research/Quick_Guides/2019-20/AustralianBushfires

[9] Celik, T., Ozkaramanli, H., & Demirel, H. (2007, April). Fire pixel classification using fuzzy logic and statistical color model. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 1, pp. I-1205). IEEE. [CrossRef]

[10] Celik, T., Demirel, H., Ozkaramanli, H., & Uyguroglu, M. (2007). Fire detection using statistical color model in video sequences. *Journal of Visual Communication and Image Representation, 18*(2), 176-185. [CrossRef]

[11] Celik, T., & Demirel, H. (2009). Fire detection in video sequences using a generic color model. *Fire safety journal, 44*(2), 147-158. [CrossRef]

[12] Yar, H., Khan, Z. A., Ullah, F. U. M., Ullah, W., & Baik, S. W. (2023). A modified YOLOv5 architecture for efficient fire detection in smart cities. *Expert Systems with Applications, 231*, 120465. [CrossRef]

[13] Yar, H., Hussain, T., Agarwal, M., Khan, Z. A., Gupta, S. K., & Baik, S. W. (2022). Optimized dual fire attention network and medium-scale fire classification benchmark. *IEEE Transactions on Image Processing, 31*, 6331-6343. [CrossRef]

[14] Tsalera, E., Papadakis, A., Voyiatzis, I., & Samarakou, M. (2023). CNN-based, contextualized, real-time fire detection in computational resource-constrained environments. *Energy Reports, 9*, 247-257. [CrossRef]

[15] Chen, T. H., Wu, P. H., & Chiou, Y. C. (2004, October). An early fire-detection method based on image processing. In *2004 International Conference on Image Processing, 2004. ICIP'04.* (Vol. 3, pp. 1707-1710). IEEE. [CrossRef]

[16] Mao, W., Wang, W., Dou, Z., & Li, Y. (2018). Fire recognition based on multi-channel convolutional neural network. *Fire technology, 54*(2), 531-554. [CrossRef]

[17] Chen, J., He, Y., & Wang, J. (2010). Multi-feature fusion based fast video flame detection. *Building and environment, 45*(5), 1113-1122. [CrossRef]

[18] Chino, D. Y., Avalhais, L. P., Rodrigues, J. F., & Traina, A. J. (2015, August). Bowfire: detection of fire in still images by integrating pixel color and texture analysis. In *2015 28th SIBGRAPI conference on graphics, patterns and images* (pp. 95-102). IEEE. [CrossRef]

[19] Cheng, G., Chen, X., Wang, C., Li, X., Xian, B., & Yu, H. (2024). Visual fire detection using deep learning: A survey. *Neurocomputing, 596*, 127975. [CrossRef]

[20] Hussain, T., Ullah, F. U. M., Khan, S. U., Ullah, A., Haroon, U., Muhammad, K., ... & de Albuquerque, V. H. C. (2022). Deep learning assists surveillance experts: Toward video data prioritization. *IEEE Transactions on Industrial Informatics, 19*(7), 7946-7956. [CrossRef]

[21] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19). [CrossRef]

[22] Khan, Z. A., Hussain, T., Ullah, F. U. M., Gupta, S. K., Lee, M. Y., & Baik, S. W. (2022). Randomly initialized CNN with densely connected stacked autoencoder for efficient fire detection. *Engineering Applications of Artificial Intelligence, 116*, 105403. [CrossRef]

[23] Yar, H., Khan, Z. A., Hussain, T., & Baik, S. W. (2024). A modified vision transformer architecture with scratch learning capabilities for effective fire detection. *Expert Systems with Applications, 252*, 123935. [CrossRef]

[24] Lee, W., Kim, S., Lee, Y. T., Lee, H. W., & Choi, M. (2017, January). Deep neural networks for wild fire detection with unmanned aerial vehicle. In *2017 IEEE international conference on consumer electronics (ICCE)* (pp. 252-253). IEEE. [CrossRef]

[25] Sharma, J., Granmo, O. C., Goodwin, M., & Fidje, J. T. (2017, August). Deep convolutional neural networks for fire detection in images. In *International conference on engineering applications of neural networks* (pp. 183-193). Cham: Springer International Publishing. [CrossRef]

[26] Dunnings, A. J., & Breckon, T. P. (2018, October). Experimentally defined convolutional neural network architecture variants for non-temporal real-time fire detection. In *2018 25th IEEE international conference on image processing (ICIP)* (pp. 1558-1562). IEEE. [CrossRef]

[27] Saponara, S., Elhanashi, A., & Gagliardi, A. (2021). Real-time video fire/smoke detection based on CNN in antifire surveillance systems. *Journal of Real-Time Image Processing, 18*(3), 889-900. [CrossRef]

[28] Muhammad, K., Ahmad, J., Mehmood, I., Rho, S., & Baik, S. W. (2018). Convolutional neural networks based fire detection in surveillance videos. *IEEE Access, 6*, 18174-18183. [CrossRef]

[29] Zhao, Y., Ma, J., Li, X., & Zhang, J. (2018). Saliency detection and deep learning-based wildfire identification in UAV imagery. *Sensors, 18*(3), 712. [CrossRef]

[30] Liu, Z. G., Yang, Y., & Ji, X. H. (2016). Flame detection algorithm based on a saliency detection technique and the uniform local binary pattern in the YCbCr color space. *Signal, Image and Video Processing, 10*(2), 277-284. [CrossRef]

[31] Wang, Z., Wang, Z., Zhang, H., & Guo, X. (2017, July). A novel fire detection approach based on CNN-SVM using tensorflow. In *International conference on intelligent computing* (pp. 682-693). Cham: Springer International Publishing. [CrossRef]

[32] Wu, X., Lu, X., & Leung, H. (2017, October). An adaptive threshold deep learning method for fire and smoke detection. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1954-1959). IEEE. [CrossRef]

[33] Maksymiv, O., Rak, T., & Peleshko, D. (2017, February). Real-time fire detection method combining AdaBoost, LBP and convolutional neural network in video sequence. In *2017 14th international conference the experience of designing and application of CAD Systems in microelectronics (CADSM)* (pp. 351-353). IEEE. [CrossRef]

[34] Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., & Baik, S. W. (2019). Efficient deep CNN-based fire detection and localization in video surveillance applications. *IEEE Transactions on Systems, Man, and Cybernetics. Systems, 49*(7), 1419-1434. [CrossRef]

[35] Hu, C., Tang, P., Jin, W., He, Z., & Li, W. (2018, July). Real-time fire detection based on deep convolutional long-recurrent networks and optical flow method. In *2018 37th Chinese control conference (CCC)* (pp. 9061-9066). IEEE. [CrossRef]

[36] Xu, H., Li, B., & Zhong, F. (2022). Light-YOLOv5: A lightweight algorithm for improved YOLOv5 in complex fire scenarios. *Applied Sciences, 12*(23), 12312. [CrossRef]

[37] Ullah, W., Ullah, A., Hussain, T., Khan, Z. A., & Baik, S. W. (2021). An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors, 21*(8), 2811. [CrossRef]

[38] Majid, S., Alenezi, F., Masood, S., Ahmad, M., Gündüz, E. S., & Polat, K. (2022). Attention based CNN model for fire detection and localization in real-world images. *Expert Systems with Applications, 189*, 116114. [CrossRef]

[39] Li, S., Yan, Q., & Liu, P. (2020). An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism. *IEEE Transactions on Image Processing, 29*, 8467-8475. [CrossRef]

[40] Sun, H., & Yao, T. (2025). Multi-Scale Construction Site Fire Detection Algorithm with Integrated Attention Mechanism. *Fire, 8*(7), 257. [CrossRef]

[41] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

[42] Ghali, R., & Akhloufi, M. A. (2024). Deep learning approach for wildland fire recognition using RGB and thermal infrared aerial image. *Fire, 7*(10), 343. [CrossRef]

[43] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520). [CrossRef]

[44] Khan, T., Khan, Z. A., & Choi, C. (2025). Enhancing real-time fire detection: An effective multi-attention network and a fire benchmark. *Neural Computing and Applications, 37*(18), 11693-11707. [CrossRef]

[45] Zhang, D., Han, S., Zhao, J., Zhang, Z., Qu, C., Ke, Y., & Chen, X. (2009, April). Image based forest fire detection using dynamic characteristics with artificial neural networks. In *2009 international joint conference on artificial intelligence* (pp. 290-293). IEEE. [CrossRef]

[46] Yao, J., Lei, J., Zhou, J., & Liu, C. (2025). FG-YOLO: an improved YOLOv8 algorithm for real-time fire and smoke detection. *Signal, Image and Video Processing, 19*(5), 346. [CrossRef]

[47] Foggia, P., Saggese, A., & Vento, M. (2015). Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion. *IEEE TRANSACTIONS on circuits and systems for video technology, 25*(9), 1545-1556. [CrossRef]



Ikram Majeed Khan received his Bachelor's degree in Software Engineering from Islamia College University, Peshawar, and a Master's degree in Computer Science from Coventry University, England, UK. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, and Visual Intelligence



Faryal Zahoor received her BS degree in Computer Science from the University of Agriculture, Peshawar, Pakistan, and her MS degree in Computer Science from Islamia College University, Peshawar. Her research interests include computer vision, machine learning, deep learning, medical image processing, and pattern recognition. She is also affiliated with the BRAINS Institute, Peshawar, where she serves as a lecturer.