RESEARCH ARTICLE

Strip Pooling Coordinate Attention with Directional Learning for Intelligent Fire Recognition in Smart Cities

Asad Ullah Haider[1,*], **Shadab Khan**[2], **Muhammad Jamal Ahmed**[3] and **Taimur Ali Khan**[4]

[1] Ilmenau University of Technology, Ilmenau 98693, Germany
[2] Department of Software Convergence, Sejong University, Seoul 05006, South Korea
[3] Departamento de Sistemas Informaticos, Universidad Politécnica de Madrid, Madrid 28031, Spain
[4] Department of IT, Saudi Media Systems, Riyadh, Saudi Arabia

Abstract

Fire detection in smart cities requires intelligent visual recognition systems capable of distinguishing fire from visually similar phenomena while maintaining real-time performance under diverse environmental conditions. Existing deep learning approaches employ attention mechanisms that aggregate spatial information isotropically, failing to capture the inherently directional characteristics of fire and smoke patterns. This paper presents DirFireNet, a novel fire detection framework that exploits directional fire dynamics through Strip Pooling Coordinate Attention (SPCA). Unlike conventional attention mechanisms, DirFireNet explicitly models vertical flame propagation and horizontal smoke dispersion via directional strip pooling operations that decompose features along horizontal and vertical axes. The framework integrates a progressive top-down fusion pathway with attention-guided weighting that synthesizes multi-scale representations from coarse to fine resolutions. Furthermore, dual global pooling captures complementary scene statistics holistic fire intensity and salient flame regions. Built upon the lightweight EfficientNetV2-S backbone, DirFireNet achieves superior accuracy while maintaining computational efficiency. Extensive experiments on the FD and BoWFire benchmark demonstrate state-of-the-art (SOTA) performance. Comprehensive ablation studies validate that directional attention contributes to accuracy gain, validating that attention mechanism provides strong inductive biases for intelligent fire recognition in smart city applications.

Keywords: fire detection, directional attention, strip pooling, smart cities, anisotropic feature learning, multi-scale fusion.

1 Introduction

Fire is one of the most destructive catastrophes due to its rapid spread and severe environmental impact. Managing fire remains a challenging task, particularly in regions with dense combustible materials such as forests, residential zones, and other sensitive environments [1, 2]. Fires may arise from human activities, equipment failures, rising temperatures, or climate change [3, 4]. Among all types, forest and

Submitted: 20 October 2025
Accepted: 26 November 2025
Published: 20 December 2025

Vol. 2, **No.** 4, 2025.
10.62762/TSCC.2025.675097

***Corresponding author:**
✉ Asad Ullah Haider
asad-ullah.haider@tu-ilmenau.de

Citation
Haider, A. U., Khan, S., Ahmed, M. J., & Khan, T. A. (2025). Strip Pooling Coordinate Attention with Directional Learning for Intelligent Fire Recognition in Smart Cities. *ICCK Transactions on Sensing, Communication, and Control*, 2(4), 263–275.

bushfires are the most hazardous because of their rapid expansion and potential to cause widespread ecological damage. For instance, the Australian bushfires in early 2020 devastated about 19 million hectares of land, destroyed over 3,000 homes, and resulted in the loss of more than 1.5 billion animals [5]. Similarly, the US Fire Administration reported over 350,000 residential fire incidents in 2021, causing thousands of fatalities and billions of dollars in losses [2]. These statistics emphasize the critical need for efficient fire detection and management strategies.

To address this issue, researchers have developed a wide range of fire detection techniques, commonly relying on visual cues or environmental sensors. Early fire recognition is essential to minimize human casualties and mitigate further damage. Traditional machine learning (ML)-based systems typically exploit features such as fire color, texture, motion, and shape [6, 7]. However, these methods face limitations, as fire appearance is highly variable due to airflow, lighting conditions, and differences in burning materials. Such factors complicate feature selection and often result in high false alarm rates or reduced accuracy.

In contrast, deep learning (DL) has significantly advanced fire detection by enabling end-to-end feature learning, which generally improves accuracy and reduces false positives. Nevertheless, most available datasets remain limited in diversity, often containing only two classes (fire and non-fire), which restricts model generalization. Moreover, complex scenarios such as sunlight resembling flames or objects that mimic fire continue to challenge even modern DL algorithms. Another obstacle lies in the deployment of these models: achieving both high precision and computational efficiency is essential for real-time applications, particularly in resource-constrained environments. Therefore, progress in fire detection requires the development of large-scale, diverse datasets that capture the complexity of real-world fire conditions. Coupled with efficient deep models, such datasets can significantly enhance the robustness and practicality of fire detection systems.

To bridge this research gap, we present DirFireNet, a novel directional attention-based framework designed specifically for intelligent fire recognition in smart cities. Our approach addresses a fundamental limitation in existing methods: conventional attention mechanisms aggregate spatial information uniformly, failing to capture the inherently anisotropic nature of fire phenomena. Physically, flames propagate vertically due to buoyancy forces, while smoke disperses along both vertical and horizontal trajectories depending on environmental airflow. DirFireNet exploits these directional characteristics through Strip Pooling Coordinate Attention (SPCA), which explicitly models horizontal and vertical fire patterns via axis-aligned feature decomposition.

1.1 Contributions

The following is a summary of the main contributions of our study:

- We propose DirFireNet, a novel fire detection framework that explicitly models the anisotropic propagation patterns of fire and smoke through Strip Pooling Coordinate Attention (SPCA). Unlike conventional isotropic attention mechanisms, SPCA decomposes spatial features along horizontal and vertical axes via directional strip pooling operations, followed by cross-dimensional interaction to capture joint directional dynamics. This physics-informed attention design enables effective discrimination of vertical flame propagation and horizontal smoke dispersion, significantly improving robustness against visually similar false positives.
- We design a progressive multi-scale feature aggregation pathway that synthesizes hierarchical representations through attention-guided top-down fusion with residual refinement. This coarse-to-fine fusion strategy enables high-level semantic guidance to resolve ambiguities in fine-grained features while maintaining gradient flow. Combined with dual global pooling that captures complementary scene statistics, holistic fire intensity via average pooling and salient flame regions via max pooling, the architecture achieves comprehensive scene understanding suitable for diverse fire scenarios.
- We conduct extensive experiments on two widely-adopted benchmarks, demonstrating that DirFireNet achieves state-of-the-art performance: on the large-scale FD dataset and on the challenging class-imbalanced BoWFire dataset, surpassing previous best methods. Comprehensive ablation studies validate that directional attention contributes substantial performance gains, confirming that modeling fire's physical propagation characteristics provides strong inductive biases for intelligent

recognition in smart city applications.

The forthcoming sections of this paper are organized as follows: Section 2 presents a comprehensive literature review, providing an overview of traditional ML and DL-based methods along with hybrid approaches. In Section 3, we elaborate on the details of our proposed methodology. Section 4 presents and discusses the datasets used for performance evaluation, parameter settings, comparative analysis, ablation study, model complexity, and detailed results. Finally, in Section 5, we conclude the paper by providing insights into potential directions for future research.

2 Related Work

Recent years have witnessed growing interest in computer vision (CV) based fire detection, aiming to automate recognition and reduce reliance on manual surveillance. These approaches can be broadly grouped into three families: conventional feature-driven methods, deep learning-based models and hybrid approaches.

2.1 Conventional Feature-Driven Methods

Earlier fire detection techniques predominantly relied on handcrafted visual cues such as color, texture, shape, and motion. Many approaches utilized RGB or YCbCr color spaces to extract fire-like regions, while fuzzy logic, statistical descriptors, and superpixel analysis were employed to improve robustness [8–11]. Optical flow has also been explored to capture motion dynamics of flames, yet its computational expense and sensitivity to lighting variations often lead to unreliable performance. Similarly, brightness-based constraints fail to capture the irregular and dynamic nature of real fire. To address these challenges, some works applied trainable classifiers such as support vector machines (SVM) using spatial-temporal covariance features [14, 15]. Despite their utility, these methods remain prone to high false alarm rates when dealing with fire-like colors, shadows, or reflective surfaces, limiting their generalizability in real-world environments.

2.2 Deep Learning-Based Models

The advent of deep learning (DL) has transformed fire detection by enabling end-to-end representation learning. Convolutional Neural Networks (CNNs) such as AlexNet, VGG, GoogLeNet, and ResNet have been widely investigated, often outperforming traditional methods in classification and detection tasks [16, 17]. For instance, ResNet50 has shown superior recognition accuracy compared to VGG16, though limited datasets restrict generalization. While plain CNN architectures offer notable improvements, their large size and computational cost pose challenges for real-time or edge deployment. To address this, lightweight variants have been explored, offering a balance between accuracy and efficiency[35]. Recent studies have also focused on model compression techniques to deploy DL models on resource-constrained devices. Approaches such as pruning redundant filters and quantizing weights to lower precision have been effective in reducing size and accelerating inference [19–21]. Hardware-aware optimizations further enhance the feasibility of real-time fire detection [18] on embedded systems.

2.3 Hybrid and Attention-Enhanced Approaches

Beyond standalone CNNs, hybrid frameworks have been proposed to integrate machine learning with deep models. For instance, recent work has proposed bilateral fusion strategies integrating transformers for global representation with CNN architectures for local feature extraction [37]. Some studies combined CNN feature extraction with SVM classifiers or fused motion analysis with CNNs for improved robustness [22, 23]. Advanced hierarchical attention frameworks employing progressive refinement strategies have demonstrated effectiveness in emphasizing overlooked regions through complementary masking and residual attention learning [33]. Similarly, region-of-interest extraction using boosting techniques (e.g., AdaBoost) has been paired with CNN-based recognition for real-time detection [24]. More recently, attention mechanisms have gained traction, enabling networks to selectively emphasize informative spatial and channel features while suppressing distractors [25–27, 38]. Such mechanisms have shown promise in addressing the challenges posed by complex fire scenes, including flame-like objects or strong illumination variations.

3 Proposed Architecture

3.1 Multi-Scale Feature Extraction

For robust fire detection across varying scales and environmental conditions, **DirFireNet** employs *EfficientNetV2-S* as the backbone network to extract hierarchical features at multiple spatial resolutions. Fire exhibits significant scale variability from small ignition points to large spreading flames while smoke can manifest as both localized plumes and diffuse atmospheric patterns. This necessitates a multi-scale representation that captures fine-grained textures,

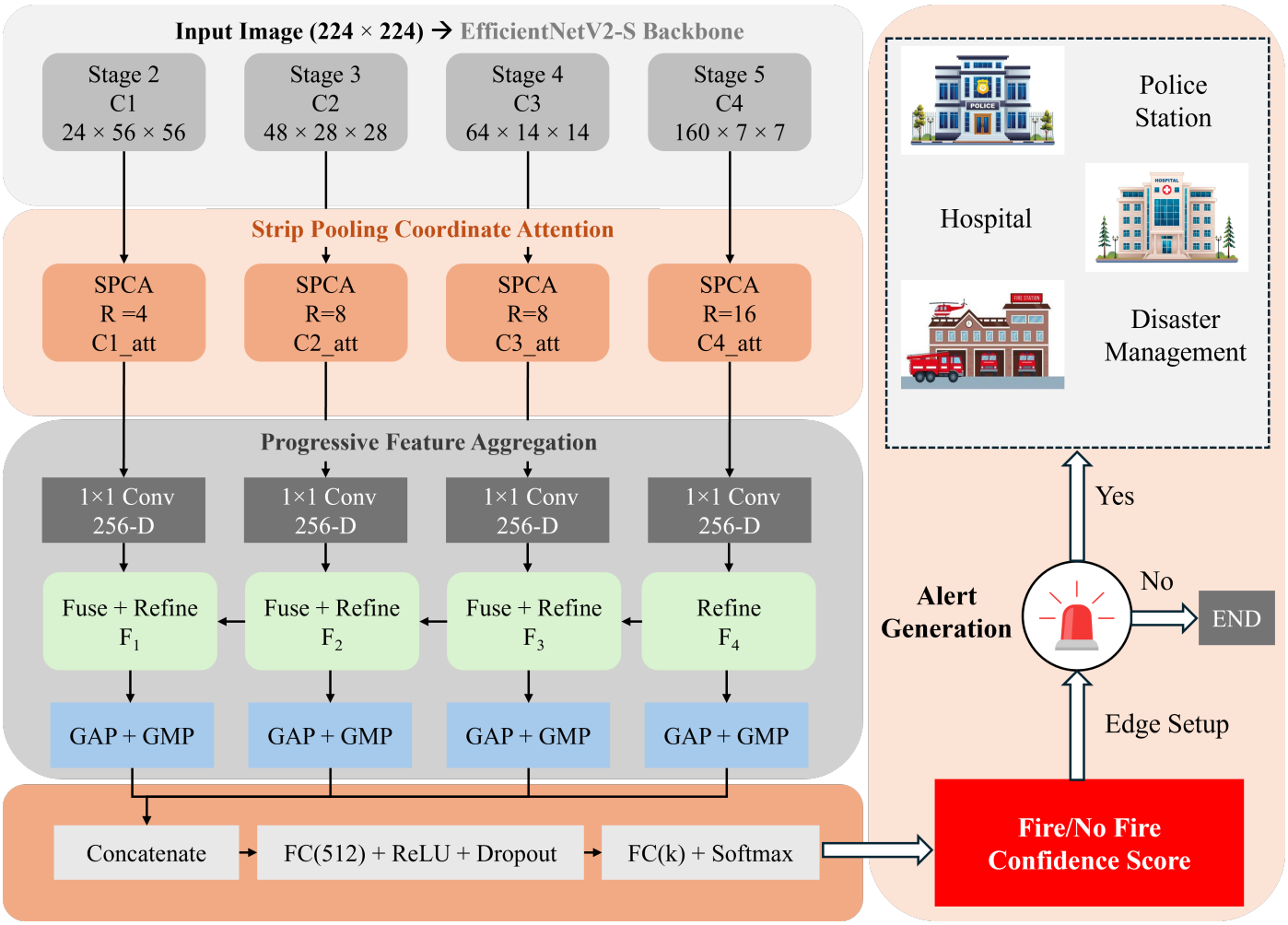

Figure 1. Architecture of DirFireNet for intelligent fire detection. Left: Deep learning pipeline comprising EfficientNetV2-S backbone, Strip Pooling Coordinate Attention (SPCA) modules for directional feature refinement, progressive multi-scale fusion with attention-guided weighting, and classification head. Right: Emergency alert system integration dispatching notifications to police stations, hospitals, and disaster management centers upon fire detection.

mid-level spatial patterns, and global contextual information simultaneously. The backbone extracts four feature maps from stages 2, 3, 4, and 5, denoted as $\{C_1, C_2, C_3, C_4\}$, with progressively increasing semantic abstraction and decreasing spatial resolution:

- **Shallow Features (C_1):** Extracted at $1/4$ input resolution with 24 channels. Preserves fine-grained spatial details essential for detecting subtle flame edges, smoke textures, and early-stage ignition points.
- **Intermediate Features (C_2 and C_3):** Captured at $1/8$ (48 channels) and $1/16$ (64 channels) resolutions, respectively. Encode mid-level semantic cues, including fire shape, intensity gradients, smoke dispersion patterns, and surrounding contextual elements (e.g., fuel sources, affected objects).
- **Deep Features (C_4):** Obtained at $1/32$ resolution with 160 channels. Provides high-level semantic abstraction, capturing scene-level understanding, fire severity classification, and global environmental context.

To enhance discriminative power and suppress irrelevant background features, each feature map is refined through SPCA Module that adaptively emphasizes fire-relevant features while attenuating noise from visually similar distractors. The resulting multi-scale feature pyramid provides a comprehensive representation that balances local fire pattern recognition with global scene understanding.

3.2 Strip Pooling Coordinate Attention for Directional Feature Refinement

Fire and smoke exhibit inherently anisotropic spatial distributions with distinct directional propagation patterns. Flames predominantly exhibit vertical upward motion due to buoyancy and convection,

while smoke disperses both vertically (rising plumes) and horizontally (lateral diffusion) depending on environmental conditions. Standard attention mechanisms such as Squeeze-and-Excitation (SE) blocks [34] and Convolutional Block Attention Module (CBAM) [36], and Coordinate Attention (CA) [39] aggregate spatial information isotropically through global average pooling or square convolutional kernels, lacking explicit modeling of these directional characteristics critical for fire detection.

To address this limitation, we propose Strip Pooling Coordinate Attention (SPCA) that explicitly models spatial dependencies along horizontal and vertical axes, enabling effective capture of spatially extended fire structures and directional smoke patterns. As illustrated in Figure 2, SPCA processes input features through directional pooling, cross-dimensional interaction, and attention-based recalibration.

3.2.1 SPCA Architecture

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, the SPCA module operates as follows:

Directional Strip Pooling: The input is decomposed into two complementary directional representations through strip pooling operations:

$$\begin{aligned} X_h &= \text{AvgPool}_{\text{height}}(X) \in \mathbb{R}^{B\times C\times 1\times W} \\ X_w &= \text{AvgPool}_{\text{width}}(X) \in \mathbb{R}^{B\times C\times H\times 1} \end{aligned} \tag{1}$$

where X_h captures horizontal spatial context by pooling along the height dimension (encoding lateral fire spread), and X_w captures vertical spatial context by pooling along the width dimension (encoding upward flame propagation).

Channel Reduction: To reduce computational overhead, both directional features undergo channel reduction via 1×1 convolutions with reduction ratio r, followed by batch normalization and ReLU activation:

$$\begin{aligned} \hat{X}_h &= \text{ReLU}(\text{BN}(\text{Conv}_{1\times 1}(X_h))) \in \mathbb{R}^{B\times C/r\times 1\times W} \\ \hat{X}_w &= \text{ReLU}(\text{BN}(\text{Conv}_{1\times 1}(X_w))) \in \mathbb{R}^{B\times C/r\times H\times 1} \end{aligned} \tag{2}$$

Cross-Dimensional Interaction: To enable information exchange between horizontal and vertical directions, the reduced features are broadcast to full spatial resolution, concatenated, processed through a 1×1 convolution to learn inter-directional correlations, and split back into separate branches:

$$\begin{aligned} \hat{X}_h^{\text{exp}} &= \text{Broadcast}(\hat{X}_h, H) \in \mathbb{R}^{B\times C/r\times H\times W} \\ \hat{X}_w^{\text{exp}} &= \text{Broadcast}(\hat{X}_w, W) \in \mathbb{R}^{B\times C/r\times H\times W} \\ X_{\text{concat}} &= \text{Concat}[\hat{X}_h^{\text{exp}}, \hat{X}_w^{\text{exp}}] \in \mathbb{R}^{B\times 2C/r\times H\times W} \\ X_{\text{inter}} &= \text{Conv}_{1\times 1}(X_{\text{concat}}) \in \mathbb{R}^{B\times 2C/r\times H\times W} \\ [X_h^{\text{inter}}, X_w^{\text{inter}}] &= \text{Split}(X_{\text{inter}}) \in \mathbb{R}^{B\times C/r\times H\times W} \end{aligned} \tag{3}$$

This cross-dimensional interaction allows the network to model complex fire patterns such as diagonal flame propagation or combined rising-spreading smoke dynamics.

Attention Map Generation: Each directional branch is expanded back to the original channel dimension through 1×1 convolutions and activated with sigmoid to produce attention maps:

$$\begin{aligned} A_h &= \sigma(\text{Conv}_{1\times 1}(X_h^{\text{inter}})) \in \mathbb{R}^{B\times C\times H\times W} \\ A_w &= \sigma(\text{Conv}_{1\times 1}(X_w^{\text{inter}})) \in \mathbb{R}^{B\times C\times H\times W} \end{aligned} \tag{4}$$

where σ denotes the sigmoid activation function, and $\text{Conv}_{1\times 1} : \mathbb{R}^{C/r} \to \mathbb{R}^{C}$ performs channel expansion.

Feature Recalibration: The attention maps are broadcast to match input dimensions and multiplicatively applied to refine the original features:

$$\text{SPCA}(X) = X \odot A_h \odot A_w \tag{5}$$

where $\odot$ denotes element-wise multiplication. This operation emphasizes spatial locations exhibiting strong directional fire/smoke characteristics while suppressing background regions, producing refined features $\text{SPCA}(X) \in \mathbb{R}^{B\times C\times H\times W}$ with enhanced directional awareness.

3.2.2 Multi-Scale Integration

SPCA is applied independently to each feature map from the EfficientNetV2-S backbone with scale-adaptive reduction ratios:

$$C_i^{\text{att}} = \text{SPCA}_{r_i}(C_i), \quad i \in \{1, 2, 3, 4\} \tag{6}$$

where $r_i \in \{4, 8, 8, 16\}$ denotes the reduction ratio for feature map C_i, balancing computational efficiency with feature expressiveness. The attention-refined features $\{C_1^{\text{att}}, C_2^{\text{att}}, C_3^{\text{att}}, C_4^{\text{att}}\}$ preserve spatial resolution while encoding directional fire patterns, providing a robust foundation for progressive feature aggregation.

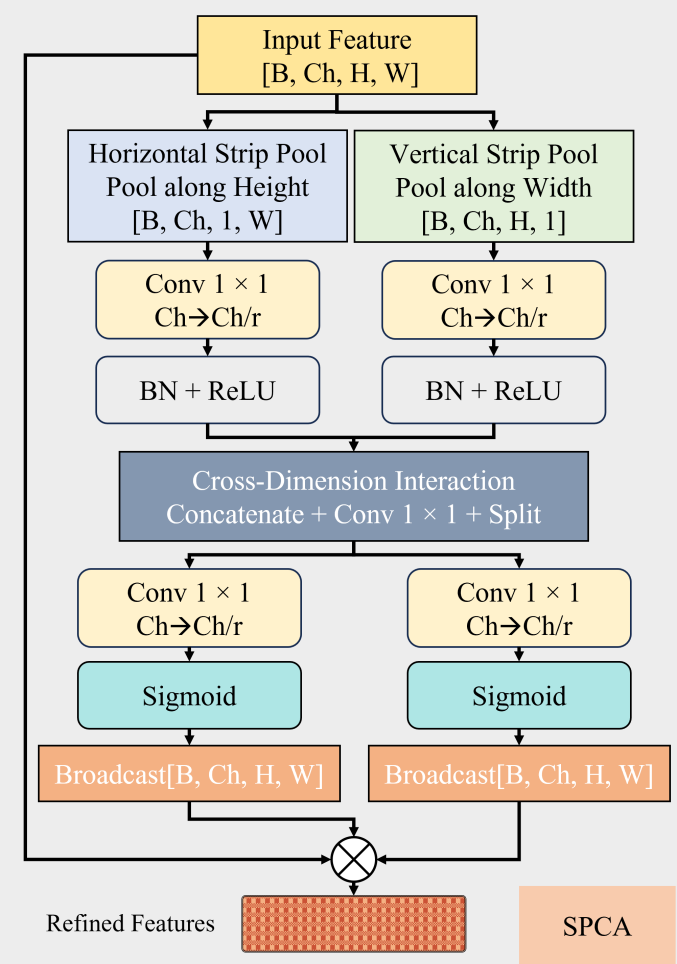

Figure 2. Architecture of the Strip Pooling Coordinate Attention (SPCA) module. The input feature undergoes directional strip pooling along horizontal (height) and vertical (width) axes.

3.3 Progressive Multi-Scale Feature Aggregation

After extracting direction-aware features through SPCA modules, the network employs a progressive top-down fusion pathway to aggregate multi-scale information, as illustrated in Figure 1. Effective fire classification requires synthesizing complementary representations across scales: shallow features capture fine-grained fire textures and flame boundaries, while deep features encode high-level semantic understanding of fire severity and scene context.

3.3.1 Top-Down Fusion Architecture

The aggregation process consists of three stages: (1) channel alignment through 1×1 convolutions, (2) progressive top-down fusion with attention-guided weighting and residual refinement, and (3) multi-scale pooling followed by classification. Starting from the deepest feature map C_4^{att} and progressively incorporating shallower layers enables high-level semantic guidance to resolve ambiguities in fine-grained features.

Channel Alignment: All SPCA-refined feature maps are projected to a unified channel dimension $D = 256$ through 1×1 convolutions:

$$\begin{aligned} &\tilde{C}_i = \text{Conv}_{1\times1}(C_i^{\text{att}}), \\ &\tilde{C}_i \in \mathbb{R}^{B\times256\times H_i\times W_i}, \quad i \in \{1,2,3,4\} \end{aligned} \tag{7}$$

This alignment facilitates effective feature fusion across different scales while maintaining computational efficiency.

Progressive Fusion and Refinement: The fusion follows a coarse-to-fine paradigm with attention-guided weighting:

$$\begin{aligned} F_4 &= \text{Refine}(\tilde{C}_4) \\ F_3 &= \text{Refine}(\text{Fuse}(\text{Upsample}(F_4), \tilde{C}_3)) \\ F_2 &= \text{Refine}(\text{Fuse}(\text{Upsample}(F_3), \tilde{C}_2)) \\ F_1 &= \text{Refine}(\text{Fuse}(\text{Upsample}(F_2), \tilde{C}_1)) \end{aligned} \tag{8}$$

where Upsample denotes bilinear interpolation to match spatial dimensions. The attention-guided fusion operation dynamically weights contributions from upsampled and lateral features:

$$\text{Fuse}(F_{\text{up}}, F_{\text{lat}}) = \alpha \odot F_{\text{up}} + (1-\alpha) \odot F_{\text{lat}} \tag{9}$$

where the spatial attention map $\alpha \in \mathbb{R}^{256\times H\times W}$ is computed as:

$$\alpha = \sigma(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(\text{Concat}[F_{\text{up}}, F_{\text{lat}}])))) \tag{10}$$

This mechanism learns to prioritize deep features in ambiguous regions (e.g., fire-like sunset colors) while emphasizing shallow features where fine-grained texture is critical (e.g., flame edges, smoke boundaries). Each fused feature undergoes residual refinement:

$$\text{Refine}(F) = F + \text{Conv}_{3\times3}(\text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(F)))) \tag{11}$$

facilitating gradient flow and adaptive feature enhancement without degrading representations.

Multi-Scale Pooling and Classification: To generate a comprehensive scene-level representation, each refined feature map undergoes dual global pooling:

$$\begin{aligned} g_i &= \text{Concat}[\text{GAP}(F_i), \\ &\text{GMP}(F_i)] \in \mathbb{R}^{512}, \quad i \in \{1,2,3,4\} \end{aligned} \tag{12}$$

where GAP (Global Average Pooling) captures holistic scene statistics and GMP (Global Max Pooling) emphasizes salient fire regions. The multi-scale descriptors are concatenated and processed through a classification head:

$$\begin{aligned} g_{\text{global}} &= \text{Concat}[g_1, g_2, g_3, g_4] \in \mathbb{R}^{2048} \\ h &= \text{Dropout}_{p=0.3}(\text{ReLU}(\text{FC}_{512}(g_{\text{global}}))) \\ y &= \text{Softmax}(\text{FC}_K(h)) \end{aligned} \tag{13}$$

where K represents the number of classes, FC_{512} denotes a fully connected layer projecting to 512 dimensions, and Dropout provides regularization. This progressive aggregation strategy synthesizes information across all spatial scales, enabling robust fire classification under diverse conditions, varying fire sizes, and complex environmental scenarios.

4 Results and Discussion

This section presents a comprehensive evaluation of the proposed fire detection framework. We begin by detailing the implementation configuration, followed by dataset descriptions and evaluation metrics. Subsequently, we analyze experimental results through quantitative comparisons with SOTA methods and ablation studies to validate the effectiveness of individual components.

4.1 Implementation Details

The proposed network is implemented using PyTorch framework. All experiments are conducted on a workstation equipped with an NVIDIA RTX 3090 GPU. The EfficientNetV2-S backbone is initialized with ImageNet-1K pre-trained weights. During training, input images are resized to 224×224 pixels and augmented with random horizontal flipping (probability 0.5), random rotation (± 15), color jittering (brightness: 0.2, contrast: 0.2, saturation: 0.2), and random Gaussian blur (kernel size: 5, σ: [0.1, 2.0]). We employ the AdamW optimizer with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate follows a cosine annealing schedule over 100 training epochs with warm-up for the first 5 epochs. The batch size is set to 16, and gradient clipping with a maximum norm of 1.0 is applied to stabilize training. Cross-entropy loss is used as the optimization objective. For regularization, dropout with rate $p = 0.3$ is applied in the classification head, and label smoothing with $\epsilon = 0.1$ is employed to prevent overconfidence.

4.2 Datasets

To comprehensively evaluate the effectiveness of our proposed method, we conduct experiments on two widely-adopted fire detection benchmarks: FD [29] and BoWFire [13]. Following standard practice, we adopt an 80:20 train-test split for both datasets. Figure 3 presents representative samples from each dataset, illustrating the diversity and challenges inherent in fire detection tasks.

FD Dataset: The FD (Fire Detection) dataset is a large-scale benchmark created by merging the comprehensive Foggia and BoWFire collections with additional fire and non-fire imagery sourced from online repositories. This extensive compilation comprises 50,000 images with balanced class distribution 25,000 fire images and 25,000 non-fire images. The dataset encompasses diverse fire scenarios including indoor fires, outdoor wildfires, industrial flames, and various environmental conditions (day, night, fog, rain), alongside challenging non-fire instances such as sunsets, artificial lights, reflections, and fire-colored objects. This substantial scale and representational depth make FD an ideal benchmark for evaluating model generalization and robustness across varied fire detection scenarios.

BoWFire Dataset: The BoWFire (Bag of Words for Fire detection) dataset represents a compact yet challenging benchmark characterized by significant class imbalance. It comprises 226 images distributed across two classes: 107 fire images and 119 non-fire images. Despite its modest size, BoWFire is valued for capturing real-world fire detection complexities, including low-resolution imagery, varying illumination conditions, and ambiguous fire-like patterns. The inherent class imbalance and limited sample size pose unique challenges for model training and generalization, making it an essential benchmark for assessing robustness under data scarcity conditions. Both datasets collectively provide complementary evaluation perspectives, FD tests large-scale generalization capability, while BoWFire evaluates robustness to limited and imbalanced data.

4.3 Evaluation Metrics

We employ a comprehensive suite of evaluation metrics following established benchmarks and SOTA fire detection methodologies [17, 29]. These metrics provide holistic assessment of model performance across different operational requirements:

- **Accuracy (A):** Measures overall classification

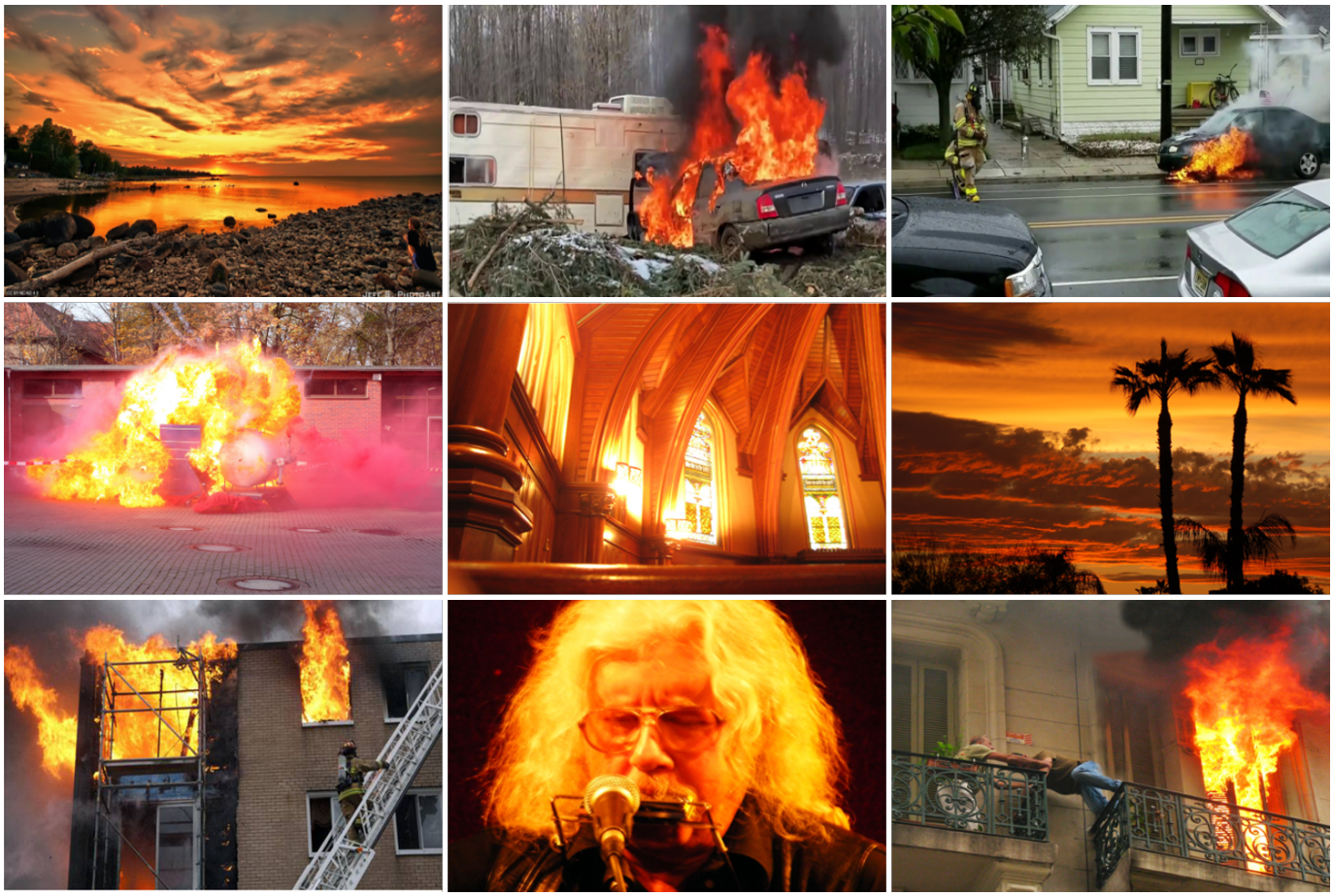

Figure 3. Representative samples from the FD and BoWFire datasets.

correctness as the ratio of correctly classified instances to total instances. While intuitive, accuracy can be misleading for imbalanced datasets.

- **Precision (P):** Quantifies the proportion of true fire detections among all fire predictions, reflecting the model's ability to minimize false alarms critical for practical deployment where false positives trigger unnecessary emergency responses.

- **Recall (R):** Represents the proportion of actual fires correctly identified, measuring the model's sensitivity to fire presence. High recall is paramount for safety applications where missing a fire (false negative) has severe consequences.

- **F1-Score (F1):** Provides the harmonic mean of precision and recall, offering a balanced assessment particularly valuable for imbalanced datasets. F1-score is essential when both false positives and false negatives carry significant costs.

These metrics are formally defined as:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{F1-Score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{14}$$

where TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) denote the counts of correctly identified fires, correctly identified non-fires, incorrectly identified fires, and missed fires, respectively.

4.4 Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of DirFireNet, we conduct comprehensive comparisons with SOTA fire detection methods on both FD and BoWFire datasets. Table 1 presents quantitative results across multiple evaluation metrics.

Table 1. Quantitative comparison of DirFireNet with state-of-the-art fire detection methods on FD and BoWFire datasets. Best results are highlighted in **bold**.

Methods	FD [29]				BoWFire [13]			
	P (%)	R (%)	F1 (%)	A (%)	P (%)	R (%)	F1 (%)	A (%)
Traditional & Early Deep Learning Methods								
FD-GCM [31]	-	-	-	-	55.00	54.00	54.00	-
FFD-ANN [32]	71.10	73.20	72.10	71.10	-	-	-	-
FPC [8]	52.00	99.90	68.40	53.90	-	-	-	-
EFD-IP [12]	75.00	15.00	25.00	-	-	-	-	-
BowFire [13]	-	-	-	-	51.00	65.00	67.00	-
CNN-based Methods								
GnetFire	88.00	98.00	92.80	92.30	79.00	93.00	85.00	84.96
ResNetFire [17]	-	-	-	-	-	-	-	92.50
ANetFire	83.30	93.20	87.90	87.20	80.00	98.00	88.00	88.05
LW-CNN [30]	82.00	81.00	81.00	81.00	86.00	78.00	77.00	79.00
CNNFire	84.60	91.30	87.90	87.30	83.00	97.00	90.00	89.82
Attention-based Methods								
EFDNet [29]	93.50	97.40	95.40	95.30	81.81	83.00	81.85	83.33
EMNFire	88.30	98.70	93.20	92.80	90.00	93.00	92.00	92.04
DFAN [28]	95.50	96.30	95.90	95.70	94.30	92.00	93.10	93.00
DirFireNet (Ours)	**96.20**	**97.10**	**96.65**	**96.50**	**95.10**	**94.50**	**94.80**	**94.70**

4.4.1 Analysis on FD Dataset

On the large-scale FD dataset, DirFireNet achieves SOTA performance across all metrics, with 96.50% accuracy, 96.65% F1-score, 96.20% precision, and 97.10% recall. Compared to the previous best method DFAN [28] (95.70% accuracy), DirFireNet demonstrates a 0.80% improvement in accuracy and 0.75% improvement in F1-score. This performance gain can be attributed to the directional attention mechanism that explicitly captures anisotropic fire patterns vertical flame propagation and horizontal smoke dispersion which conventional isotropic attention methods fail to model effectively.

The high precision (96.20%) indicates DirFireNet's robustness in minimizing false positives, crucial for reducing unnecessary emergency alerts in smart city deployments. Meanwhile, the excellent recall (97.10%) demonstrates the network's sensitivity in detecting actual fire instances, which is paramount for safety-critical applications where missing a fire event has severe consequences. Notably, DirFireNet outperforms traditional methods by substantial margins (e.g., 25.4% accuracy improvement over FFD-ANN), validating the effectiveness of deep learning and attention mechanisms for fire detection.

4.4.2 Analysis on BoWFire Dataset

On the challenging BoWFire dataset characterized by limited samples (226 images) and class imbalance, DirFireNet achieves 94.70% accuracy and 94.80% F1-score, surpasses DFAN by 1.70% in accuracy and 1.70% in F1-score. This performance gain is particularly significant given the dataset's constraints, demonstrating DirFireNet's superior generalization capability under data scarcity conditions.

The balanced precision (95.10%) and recall (94.50%) indicate that DirFireNet effectively handles class imbalance without sacrificing either metric. Traditional methods struggle on BoWFire due to limited training samples, FD-GCM and the BowFire method achieves only 54-67% F1-scores. Even recent attention-based methods like EFDNet show relatively weak performance (81.85% F1-score), highlighting the challenge of this dataset. DirFireNet's substantial improvement (12.95% F1-score gain over EFDNet) underscores the effectiveness of directional strip pooling attention in learning discriminative features from limited data by explicitly modeling fire's physical propagation patterns.

DirFireNet consistently outperforms methods using standard attention (EFDNet, DFAN), validating the importance of modeling directional fire characteristics. The performance gap between DirFireNet and baselines is larger on BoWFire (1.70% improvement) than FD (0.80% improvement), suggesting that directional priors are particularly valuable when training data is limited. DirFireNet maintains balanced

Table 2. Ablation study showing the contribution of each component in DirFireNet. SPCA: Strip Pooling Coordinate Attention; PFA: Progressive Fusion with Attention; DP: Dual Pooling (GAP+GMP).

Components				FD		BoWFire	
Baseline	SPCA	PFA	DP	F1 (%)	A (%)	F1 (%)	A (%)
✓				92.30	92.10	88.50	88.20
✓	✓			94.80	94.60	91.40	91.10
✓	✓	✓		95.90	95.70	93.20	93.00
✓	✓	✓	✓	**96.65**	**96.50**	**94.80**	**94.70**

precision-recall trade-offs across both datasets, unlike some methods (e.g., FPC achieves 99.90% recall but only 52.00% precision), making it suitable for practical deployment where both false positives and false negatives carry costs. These results demonstrate that DirFireNet establishes new SOTA performance on both benchmarks, validating the effectiveness of directional strip pooling coordinate attention for intelligent fire recognition.

4.5 Ablation Studies

To validate the contribution of individual components in DirFireNet, we conduct comprehensive ablation studies on both datasets. Table 2 presents the results of systematically adding components to a baseline architecture.

4.5.1 Baseline Architecture

The baseline model consists of EfficientNetV2-S backbone with standard channel alignment (1×1 convolutions to 256-D), simple element-wise addition for multi-scale fusion, and single global average pooling before classification. This baseline achieves 92.10% accuracy on FD and 88.20% accuracy on BoWFire, providing a strong foundation but lacking specialized components for fire detection.

4.5.2 Effect of Strip Pooling Coordinate Attention (SPCA)

Integrating SPCA modules after each backbone stage yields substantial improvements: +2.50% accuracy on FD and +2.90% accuracy on BoWFire. This demonstrates that directional attention effectively captures anisotropic fire patterns. SPCA's explicit modeling of horizontal and vertical spatial dependencies enables the network to distinguish directional smoke dispersion and vertical flame propagation from isotropic background patterns. The larger improvement on BoWFire suggests that directional priors are particularly valuable when training data is limited, as SPCA provides strong inductive bias aligned with fire physics.

4.5.3 Effect of Progressive Fusion with Attention (PFA)

Adding attention-guided progressive fusion on top of SPCA brings additional gains: +1.10% on FD and +1.90% on BoWFire. The top-down fusion pathway enables high-level semantic information to guide lower-level feature refinement, helping resolve ambiguities between fire and fire-like objects (sunsets, lights). The learnable spatial attention weights in the fusion module adaptively balance contributions from different scales based on spatial context prioritizing deep features in ambiguous regions and shallow features where fine-grained texture is critical. The larger improvement on BoWFire again indicates that sophisticated fusion strategies help maximize information extraction from limited data.

4.5.4 Effect of Dual Pooling (GAP + GMP)

Combining global average pooling and global max pooling provides complementary information, yielding final improvements of +0.75% on both datasets. While GAP captures holistic scene statistics (overall fire intensity, smoke coverage), GMP emphasizes salient local peaks (intense flame regions, fire hotspots). This dual pooling strategy ensures the network leverages both diffuse fire characteristics (captured by GAP) and localized high-intensity responses (captured by GMP), resulting in more robust classification decisions.

4.5.5 Component Interaction Analysis

Table 3. Component interaction analysis showing individual and combined contributions. Results on FD dataset (Accuracy %).

Configuration	Accuracy (%)
Baseline	92.10
Baseline + SPCA only	94.60
Baseline + PFA only	93.80
Baseline + DP only	92.90
Baseline + SPCA + PFA	95.70
Baseline + SPCA + DP	95.30
Baseline + PFA + DP	94.50
Full DirFireNet (All)	**96.50**

Table 3 analyzes component interactions. SPCA provides the largest individual contribution (+2.50%), validating directional attention as the core innovation. PFA and DP offer smaller individual gains (+1.70% and +0.80%), but their combination with SPCA yields synergistic effects the full model (96.50%) outperforms the sum of individual improvements, indicating that components complement each other effectively.

4.5.6 Directional Attention Decomposition

Table 4. Analysis of directional attention components in SPCA. Results on FD dataset.

SPCA Variant	F1 (%)	A (%)
Baseline (no attention)	92.30	92.10
Horizontal pooling only	93.80	93.60
Vertical pooling only	93.50	93.30
H + V (no interaction)	94.30	94.10
Full SPCA (with cross-interaction)	**94.80**	**94.60**

Table 4 decomposes SPCA to analyze directional components. Horizontal pooling (+1.50%) slightly outperforms vertical pooling (+1.20%), suggesting that horizontal smoke dispersion patterns are marginally more discriminative than vertical flame propagation in the tested datasets. However, combining both directions without cross-dimensional interaction yields +2.00% improvement, and adding the interaction module provides an additional +0.50% gain (94.60% vs 94.10%), demonstrating that learning correlations between horizontal and vertical patterns (e.g., diagonal flame propagation, combined rising-spreading smoke) further enhances performance.

5 Conclusion

In this paper, we presented DirFireNet, a novel fire detection framework that exploits the directional characteristics of fire and smoke through Strip Pooling Coordinate Attention (SPCA). Unlike conventional attention mechanisms that aggregate spatial information isotropically, DirFireNet explicitly models vertical flame propagation and horizontal smoke dispersion through directional strip pooling operations, combined with progressive multi-scale fusion. Extensive experiments on FD and BoWFire benchmarks demonstrate that DirFireNet achieves state-of-the-art performance with 96.50% and 94.70% accuracy respectively, surpassing previous best methods. Comprehensive ablation studies validate the contribution of each component, with SPCA providing the largest individual performance gain, confirming that directional attention modeling of fire's physical propagation patterns is the core innovation. Future work will explore multi-modal sensor fusion, temporal modeling for video-based detection, and real-world deployment validation in smart city environments.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Taimur Ali Khan is affiliated with the Department of Information Technology, Saudi Media Systems, Riyadh, Saudi Arabia. The authors declare that this affiliation had no influence on the study design, data collection, analysis, interpretation, or the decision to publish, and that no other competing interests exist.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Urza, A. K., Hanberry, B. B., & Jain, T. B. (2023). Landscape-scale fuel treatment effectiveness: lessons learned from wildland fire case studies in forests of the western United States and Great Lakes region. *Fire Ecology, 19*(1), 1. [CrossRef]

[2] Tan, C., & Feng, Z. (2023). Mapping forest fire risk zones using machine learning algorithms in Hunan province, China. *Sustainability, 15*(7), 6292. [CrossRef]

[3] *Residential fire estimate summaries (2014-2023)*. (n.d.). U.S. Fire Administration. Available from: https://www.usfa.fema.gov/statistics/residential-fires/

[4] Keith, D. A., Allen, S. P., Gallagher, R. V., Mackenzie, B. D., Auld, T. D., Barrett, S., ... & Tozer, M. G. (2022). Fire-related threats and transformational change in Australian ecosystems. *Global Ecology and Biogeography, 31*(10), 2070-2084. [CrossRef]

[5] Gaur, A., Singh, A., Kumar, A., Kumar, A., & Kapoor, K. (2020). Video flame and smoke based fire detection algorithms: A literature review. *Fire technology, 56*(5), 1943-1980. [CrossRef]

[6] Swain, D. L., Abatzoglou, J. T., Kolden, C., Shive, K., Kalashnikov, D. A., Singh, D., & Smith, E. (2023). Climate change is narrowing and shifting prescribed fire windows in western United States. *Communications Earth & Environment, 4*(1), 340. [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] 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]

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

[10] Frizzi, S., Kaabi, R., Bouchouicha, M., Ginoux, J. M., Moreau, E., & Fnaiech, F. (2016, October). Convolutional neural network for video fire and smoke detection. In *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society* (pp. 877-882). IEEE. [CrossRef]

[11] 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]

[12] 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]

[13] 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]

[14] Habiboğlu, Y. H., Günay, O., & Çetin, A. E. (2012). Covariance matrix-based fire and flame detection method in video. *Machine Vision and Applications, 23*(6), 1103-1113. [CrossRef]

[15] Ko, B. C., Cheong, K. H., & Nam, J. Y. (2009). Fire detection based on vision sensor and support vector machines. *Fire Safety Journal, 44*(3), 322-329. [CrossRef]

[16] 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]

[17] 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]

[18] 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]

[19] Ye, S., Feng, X., Zhang, T., Ma, X., Lin, S., Li, Z., ... & Wang, Y. (2019). Progressive dnn compression: A key to achieve ultra-high weight pruning and quantization rates using admm. *arXiv preprint arXiv:1903.09769*.

[20] Carreira-Perpinan, M. A., & Idelbayev, Y. (2018, June). " Learning-Compression" Algorithms for Neural Net Pruning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8532-8541). IEEE. [CrossRef]

[21] Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016, September). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision* (pp. 525-542). Cham: Springer International Publishing. [CrossRef]

[22] 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]

[23] 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]

[24] 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]

[25] Shen, C., Qi, G. J., Jiang, R., Jin, Z., Yong, H., Chen, Y., & Hua, X. S. (2018). Sharp attention network via adaptive sampling for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology, 29*(10), 3016-3027. [CrossRef]

[26] 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]

[27] 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]

[28] 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]

[29] 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]

[30] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020, June). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11531-11539). IEEE. [CrossRef]

[31] 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]

[32] 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]

[33] Jia, Y., Zeng, Y., & Guo, H. (2025). Cascade Aggregation Network for Accurate Polyp Segmentation. *IET Systems Biology, 19*(1), e70036. [CrossRef]

[34] Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(8), 2011-2023. [CrossRef]

[35] Usman, M. T., Khan, H., Rida, I., & Koo, J. (2025). Lightweight transformer-driven multi-scale trapezoidal attention network for saliency detection. *Engineering Applications of Artificial Intelligence, 155*, 110917. [CrossRef]

[36] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision* (pp. 3-19). Cham: Springer International Publishing. [CrossRef]

[37] Khan, H., Usman, M. T., & Koo, J. (2025). Bilateral feature fusion with hexagonal attention for robust saliency detection under uncertain environments. *Information Fusion, 121*, 103165. [CrossRef]

[38] Khan, H., Usman, M. T., Rida, I., & Koo, J. (2024). Attention enhanced machine instinctive vision with human-inspired saliency detection. *Image and Vision Computing, 152*, 105308. [CrossRef]

[39] Hou, Q., Zhou, D., & Feng, J. (2021, June). Coordinate Attention for Efficient Mobile Network Design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13708-13717). IEEE. [CrossRef]

Asad Ullah Haider is a master's student at Ilmenau University of Technology, Germany, specializing in the fields of Machine Learning, Deep Learning, and Artificial Intelligence. His academic work focuses on exploring intelligent algorithms, data-driven models, and advanced computational techniques to solve real-world problems. He is actively engaged in research and projects that involve the application of AI to various domains, aiming to contribute to the development of efficient and innovative solutions.

Shadab Khan received the Bachelor's degree in Computer Science from Islamia College Peshawar, Peshawar, Pakistan. He is currently pursuing the Master's degree with the Department of Software Convergence, Sejong University, Seoul, South Korea. He is also working as a Research Assistant at Sejong University. His major research interests include object detection, and segmentation for effective disaster management, as well as various sub-domains of machine learning, deep learning, and computer vision for real-world applications.

Muhammad Jamal Ahmed received his bachelor's degree in Computer Science and IT from University of Engineering and Technology, Peshawar, Pakistan, in 2016 and then pursued his M.Sc. in Computing Science and Engineering from Kyungpook National University, Daegu, South Korea. He is currently working as an early-stage researcher in the Department of Informatics, Universidad Politécnica de Madrid, Spain. His research interests include Artificial Intelligence, Deep Learning, and Time Series Analysis

Taimur Ali Khan holds a Bachelor's degree in Information Technology from the University of Agriculture, Peshawar. He is currently working as a Senior Developer and IT Consultant at Saudi Media Systems. With extensive experience in software development and IT solutions, he integrates academic knowledge with real-world applications. His research interests span machine learning, deep learning, and their applications in intelligent information systems, as well as system architecture, enterprise software development, and emerging technologies. He aims to bridge practical industry expertise with cutting-edge research to develop innovative and scalable AI-driven solutions.