Scientific Reports

https://doi.org/10.1038/s41598-026-64460-5

Article in Press

Scene-aware attention network and medium-scale drone dataset for diverse fire classification

Received: 14 March 2026

Accepted: 25 July 2026

Published online: 22 August 2026

Cite this article as: Danish S., Khan S.U., Minh Dang L. *et al.* Scene-aware attention network and medium-scale drone dataset for diverse fire classification. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-64460-5

Sufyan Danish, Samee Ullah Khan, L. Minh Dang, Hyoung-Kyu Song & Hyeonjoon Moon

Scene-Aware Attention Network and Medium-Scale Drone Dataset for Diverse Fire Classification

Sufyan Danish[1], **Samee Ullah Khan**[2], **L. Minh Dang**[3,4], **Hyoung-Kyu Song**[4], **and Hyeonjoon Moon**[1,*]

[1]Department of Computer Science and Engineering, Sejong University, Seoul, 05006, Republic of Korea
[2]Advanced Research and Innovation Center (ARIC), Khalifa University of Science and Technology, Abu Dhabi, UAE.
[3]Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[4]Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul, 05006, Republic of Korea
[*]corresponding: hmoon@sejong.ac.kr

ABSTRACT

Vision-based fire detection is a critical task for ensuring urban safety and protecting buildings against potential hazards. Although mainstream approaches practice conventional attention modules with deep models that are well-suited for static fire environments, they have limited outcomes in their diverse nature. These attention modules are mainly composed of conventional convolutional layers that generate redundant and overlapped spatial patterns in the case of complex fire scenarios, especially remote sensing views. To overcome these challenges, an intelligent vision model is desirable which has the capabilities to differentiate complex fire scenes with contextual information, particularly from remote sensing viewpoints. To address these challenges, an Adaptive Fire Network (AFN) is proposed that efficiently discriminates diverse fire scenes, geographical location and environmental conditions. To obtain more informative spatial features, a lightweight ConvNeXtBase is integrated with the Scene-Aware Attention Network (SAAN). Moreover, SAAN is mainly composed of the Adaptive Channel Module (ACM) and the Adaptive Spatial Module (ASM) to effectively capture global dependencies across channels, followed by local spatial patterns associated with fire. In the SAAN, depth-wise separable convolution is utilized which allows the model to gather more enhanced features from the input tensor. Additionally, the channel-wise processing leads to more understanding of the input features, potentially enhancing the model's ability to make accurate predictions. Extensive experiments were conducted on three challenging fire datasets, Foggia, FireNet, and the newly developed Drone Fire Vision (DFV), achieving accuracies of 99.96%, 98.97%, and 90.78%, respectively. The results demonstrate that AFN significantly enhanced existing state-of-the-art methods, advancing fire detection and classification capabilities in dynamic and remote sensing environments.

1 Introduction

Disaster management has gained significant attention in numerous fields such as computer vision, soft computing, and environmental science, as these technologies play a crucial role in enhancing the efficiency and effectiveness of fire detection systems, especially in Remote Sensing applications. It is broadly categorized into natural and technological disasters that involve nuclear plant accidents, terrorism, and hazardous materials while others contain floods, earthquakes, and fire[1]. Fire disasters are extremely hazardous and occur due to natural causes, environmental effects, or human-made errors. Once exploded, it spread rapidly on a large scale, thereby causing severe damage to the properties or building structures, hundreds of human lives, and also largely affecting the environment and atmosphere for plants, forests, and animals[2]. In recent times, due to climate changes, several large-scale hazardous fires have exploded across different forests, urban, and industrial areas. As reported by the International Association of Fire and Rescue Services[3] In the year 2020, there are 3.3 million hazardous wildfire incidents occurred on average across the globe in 48 different countries of the world. As a result of these hazardous fire incidents, approximately 42,000 people lost their lives in the year 2020. In 2025, a devastating fire in Los Angeles was reported that caused extensive destruction over 5,316 structures and affecting an area of about 13,750 hectares[4]. Furthermore, the number of fire incidents in China has reduced from 60,000 in 1983 to an average of about 2000 forest fires in 2022[5]. Early fire detection plays a vigorous role in the protection of the general public and in reducing damage to properties. However, the current fire-related research mainly used fire videos/images captured from static vision sensors while a limited study focuses on fire detection using drone datasets. However, the main problem is that there is no publicly available data for drones or with certain data samples. Therefore, this research's fundamental objective is to develop a drone dataset composed of various types of visual data and generalized vision models compatible with fire surveillance. We frame our work as a trust-enhancing approach for remote sensing applications, emphasizing reliability, which includes accurate alerts with a low false positive rate

(FPR) and false negative rate (FNR), robustness, which involves performance across diverse conditions and viewpoints, and transparency, which refers to readable attention responses, all aimed at proactive risk reduction.

In the past decades, several sensors-based techniques have been utilized for fire detection and classification to protect human lives and property early from harmful fire incidents. To this end, numerous studies[6–8] have employed various sensors to detect fire at an early stage. However, it detects smoke[7], measures the surrounding temperature or heat[6] and use light indicators to sense fire flames[8]. Furthermore, it is low-cost and informal to operate, but certain downsides: Firstly, these sensors require closeness to the fire, thereby having a limited range and can only be used for indoor fire detection. Secondly, with the passage of time, the sensitivity of these sensors also degrades, because it not only requires a longer time to detect fire and smoke but also generates false triggering alarms in case of other similar situations[9].

To tackle the drawbacks of false triggering and the limitation of a wide coverage area, these days researchers focus on vision-based sensors because it is more reliable than environmental sensors and respond very quickly[10]. In addition, with the advancement in sensor technology, surveillance cameras have been widely installed across various public places such as parks, airports, subways, office buildings, hospitals, shopping malls, and in many residential areas. For vision-based sensors, the researchers followed different approaches for fire detection and classification. Initially, Traditional Machine Learning (TML) was employed mainly concentrating on color features, shape, motion and texture[11]. Moreover, it relies on handcrafted features, but identifying the best features is difficult due to various factors such as lighting, fire shape, and airflow. It is still challenging to find the right balance between accuracy, loss, and False Positive Rate (FPR). To address such difficulties, researchers diverted to advanced Deep Learning (DL) fire detection and classification techniques[12]. DL models are famous for classification and recognition due to their high performance in different domains[10]. DL-based approaches show high performance in fire detection within surveillance systems, substantially reducing false alarm rates compared to TML techniques. Nevertheless, these approaches demonstrate limited abilities in accurate classification and localization of challenging scenes of fire encompassing things that resemble fire, lighting that simulates fire, and scenes where sunlight imitates flames. Several fire-related studies and developed various techniques but have certain drawbacks and partially covered the existing challenges. A brief explanation of these trials is provided in the subsequent paragraph.

1.1 Major Challenges

A review of existing fire detection literature reveals that researchers commonly employ attention modules, such as bottlenecks combining channel and positional attentions. However, these models primarily focus on high-level pixel features while neglecting low-level details that are critical for classification. Since fire is highly dynamic and lacks fixed patterns, such attentions often produce redundant feature maps and overlapped spatial patterns, degrading performance. Furthermore, bottleneck-based dimensionality reduction discards fine-grained details essential for accurate detection, particularly in drone surveillance. Simplified representations hinder the capture of diverse fire patterns in complex scenarios, leading to non-negligible false positives/negatives.

Conventional surveillance cameras offer limited coverage and incur high costs, motivating the use of UAVs for fire detection[13,14]. UAVs provide flexible, wide-area sensing suitable for municipal and remote sensing applications, initially employed for navigation and video capture in forested and residential areas[15]. With technological advances, commercial UAVs have become accessible, enabling efficient fire monitoring with benefits such as large-area coverage, day-and-night operation, cost-effectiveness, environmental adaptability, and the ability to carry diverse payloads[16]. These advantages underscore the demand for reliable, and interpretable vision models aligned with trustworthiness requirements in remote sensing.

This study addresses these challenges by proposing the Adaptive Fire Network (AFN), which integrates a ConvNeXtBase backbone with a Scene-Aware Attention Network (SAAN). SAAN comprises Adaptive Channel (ACM) and Adaptive Spatial (ASM) modules that capture both global dependencies and local spatial patterns using depth-wise separable convolutions for efficient feature extraction. By incorporating both high- and low-level pixel information, AFN produces detailed representations, enhancing adaptability across diverse fire scenarios and environments. A UAV-based implementation further extends its applicability to enclosed and open settings. The major contributions of this study are summarized below:

- We propose the Adaptive Fire Network (AFN), a novel framework for reliable fire classification in dynamic environments. By intelligently analyzing fine-grained patterns and contextual cues, particularly in remote-sensing scenarios, AFN ensures accurate fire detection, facilitating early detection and supporting timely disaster management to safeguard infrastructure and human lives.

- The Scene-Aware Attention Network (SAAN) introduces an innovative dual-depth attention mechanism that jointly captures global and local contextual information. By integrating multi-scale features and adaptively focusing on critical regions, SAAN enables robust classification of dynamic fire scenes under diverse conditions. Its real-time performance strengthens decision-making for disaster mitigation and enhances public safety.

- To address the lack of drone-specific fire datasets, we curated and introduced the Drone Fire Vision (DFV) dataset, which will be publicly released for the research community. Extensive experiments on Foggia's, FireNet, and DFV demonstrate AFN's robustness across heterogeneous environments and sensor modalities. Detailed analysis against multiple backbones and attention modules further highlights its effectiveness for real-world deployment.

The remaining paper is structured in the following way: the literature review is discussed in **Section 2**, while the technical information about the proposed framework is provided in **Section 3**. Moreover, the details about dataset, evaluation metrics, implementation detail and empirical results are given in **Section 4**. The performance of the proposed model is compared with existing methods in **Section 5**. At the end, the research is concluded with future direction in **Section 7**.

2 Related Work

Fire detection has been extensively studied over the years, with early methods primarily relying on sensor-based technology. These sensors are mainly employed for smoke[7] and gas detection[17] evaluate the intensity of the smoke particles in the atmosphere while others[18] have been utilized for heat and smoke detection. Although sensor-based fire systems are effective in detecting and alarming fires but face several challenges in the long term. These challenges include sensor sensitivity decline, delayed smoke and heat detection, sensor placement limitations, and high false alarm rates. The collective implications of these factors significantly reduce their reliability in the long term. In the United Kingdom (UK), about £1 billion is lost every year due to the false triggering of fire alarms[9].

To address the above issues, researchers have recently focused on vision-based systems for fire detection. These fire detection approaches are typically developed either through the ML approaches that manually engineered features[19] or via end-to-end DL approaches[10]. To embark, recent research endeavors are deeply reviewed and decided to be categorized into two sections i.e. ML and DL where the brief descriptions are presented in the following subsections.

2.1 Machine Learning Based Fire Systems

ML-based fire detection methods focused on extracting discriminative handcrafted features to classify images into their respective categories. These are low-level features that include color, texture, and fire motion features. Fire detection based on color is one of the preliminary techniques used to detect fires in ongoing scenes. In this technique, an image series is used to validate channels of color that closely resemble fire. The prior study contains different color-based approaches for this purpose[11]. These methods used statistical color models followed by background removal mechanisms developed to identify fire pixels. Moreover, the proposed technique[20] utilizes the HSV color space with enforced constraints for detecting pixels of fire in the source image. Real-time fire detection over edge devices is more reliable and easier to deploy using color-based feature extraction models. However, this technique has significant challenges involving a high false alarm rate and fails in fire-like scenes[21]. The main factor is the rapid environmental change that affects of the model performance and delays accurate fire detection. To address these challenges numerous struggles have been undertaken to minimize the false rate and enhance the performance. Some researchers integrate color characteristics with motion to enable early fire detection. These methods mitigate false 23 alarms, although their limited accuracy remains a challenge[21]. Furthermore, handcrafted feature extraction presents a significant challenge in fire detection due to the variation in fire attributes such as structure, color, and lighting conditions. Therefore, several researchers shifted their attention toward end-to-end DL models for fire detection and classification, where a brief explanation is covered in the next subsection.

2.2 Deep Learning-Based Fire Systems

To address the challenges of ML algorithms, researchers' attention is diverted toward Convolutional Neural Network (CNN) models for complex fire analyses. The automatic feature learning capabilities of CNN-based models make them more reliable. Different deep learning-based methods have been used for fire detection and classification, such as[22–25, 25–29]. Furthermore, a study presented in[30] applied various pre-trained models such as AlexNet, GoogleNet, and VGG, where each has diverse feature learning capabilities. Their study found that Google Net was the most effective and able to extract more significant information from the given input image. Similarly, in[31], a CNN is trained for binary classification with a massive number of kernels with variant sizes to analyze fire from diverse filtering techniques. However, due to complex fire scenes in challenging environments, researchers are then focused on hybrid models by integrating DL and ML for more accurate discrimination of fire. For instance, a hybrid model[32] is proposed by integrating a Support Vector Machine and CNN to improve the precision score on testing data. The model used a Haar feature and an AdaBoost cascade classifier initially, followed by four CNN layers for extracting predominant patterns. Later, a two-tier SVM approach was utilized for classifying fire scenes. Similarly, in[33] features are obtained via CNN and later employed PCA to select more robust features and dimensionality reduction to minimize the time complexity when deployed in real-time. Despite these advancements, existing deep learning methods still face challenges in handling complex fire scenarios, including variations in scale, illumination, occlusion, and background

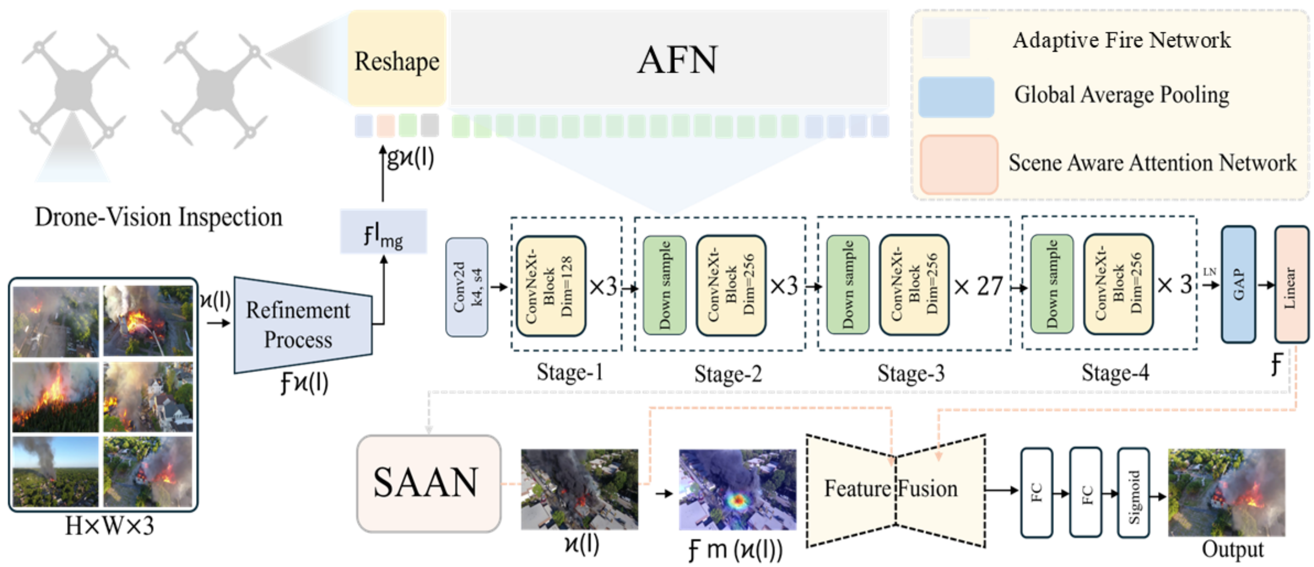

Figure 1. Detailed architecture of the proposed AFN framework. The input image is resized to 224×224×3 and processed through the ConvNeXtBase backbone. The backbone generates hierarchical feature representations at progressively reduced spatial resolutions. The final feature tensor is passed to the Scene-Aware Attention Network (SAAN), which consists of the Adaptive Channel Module (ACM) and Adaptive Spatial Module (ASM). The refined features are then pooled and forwarded to the classifier for binary fire/non-fire prediction.

similarity. Furthermore, in the literature, fire-related visual intelligence tasks are generally categorized into two main directions: (i) fire detection, which focuses on localizing fire regions using bounding box-based frameworks such as YOLO variants, and (ii) fire scene classification, which focuses on assigning global semantic labels (fire or non-fire, or multi-class fire categories) at the image level. Although YOLO-based detectors are widely adopted in UAV-based fire monitoring systems, they are primarily designed for spatial localization and are not directly comparable to image-level classification frameworks. In contrast, this study focuses on fire scene classification, where the objective is to learn discriminative global representations of fire patterns under diverse environmental conditions rather than performing object localization. The DL models performed well as compared to the ML approaches, but still, some gaps have not been considered yet as mentioned in Sub-section 1.1. To overcome such difficulties, numerous attention models are widely used that mainly concentrate on foreground objects and partially element the background information.

2.3 Attention-Based Fire Systems

Several computer vision based researchers explored attention-based models for fire detection and classification[34]. Attention mechanisms have demonstrated remarkable improvements in various computer vision tasks, encompassing recognition and detection. In the study[2], utilized transfer learning to develop a lightweight model for fire detection and an attention-based model was used to detect fire in adverse weather conditions. Similarly, in[35] introduced the Dual Semantic Attention (DSA) module, which employs attention mechanisms within convolutional kernels to better capture semantic information during convolutional operations. This module dynamically chooses, and merges feature maps from different kernel scales, enhancing the model's ability to discern relevant information. The authors incorporated the DSA module into the ResNet50 architecture, creating the DSA-Residual module, thus improving its performance. Furthermore,[36] refined attention learning for fire models by incorporating dual attention, where initially, the most pertinent channels are accentuated, followed by the application of modified spatial attention on the important channels to capture intricate pixel-wise spatial details differentiating fire from non-fire regions. In summary, there are still various limitations in current attention techniques for complex fire detection as explained in the subsection 1.1.

3 The Proposed Framework

In this section, the technical information of the modules is applied to the complex fire scene classification. For the training process, a comprehensive analysis is performed where various pre-trained CNN-based models are utilized. After an extensive experimental process, a ConvNeXt-B architecture is selected as an optimal backbone model followed by AFN based on their

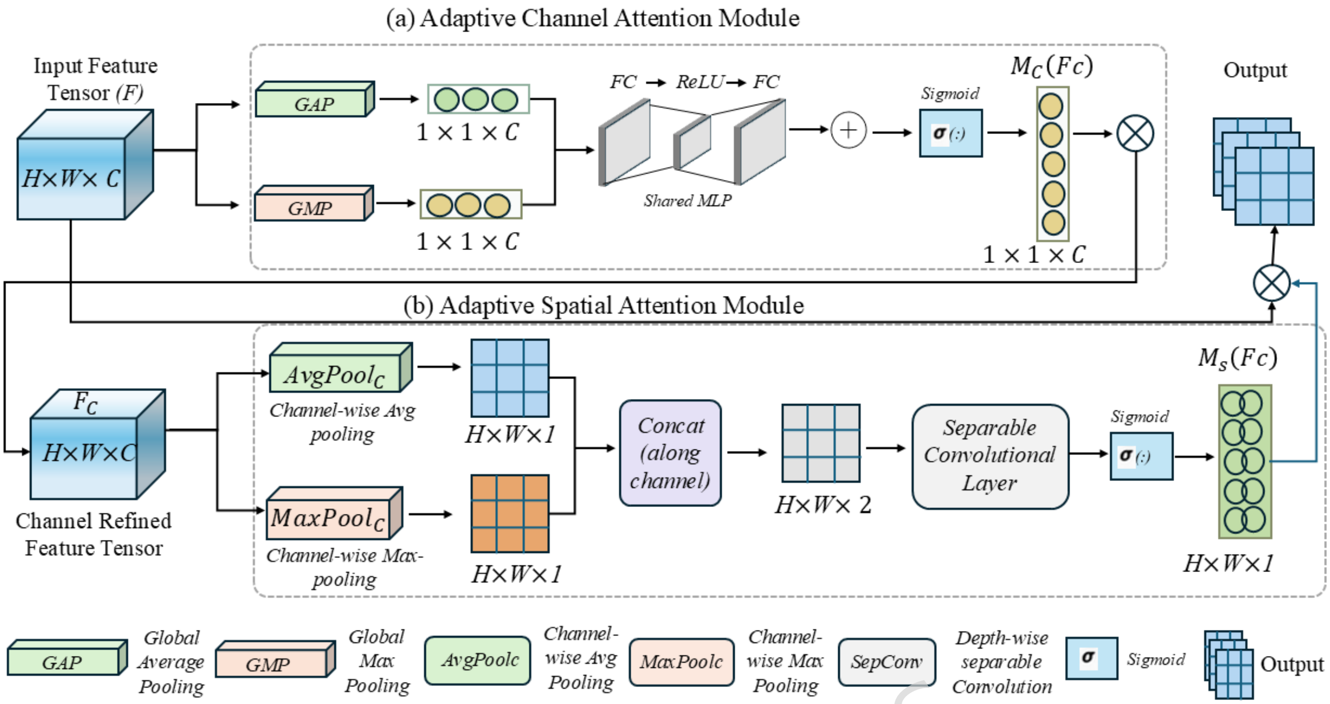

Figure 2. Architecture of the proposed Scene-Aware Attention Network (SAAN), consisting of (a) Adaptive Channel Attention Module (ACM) and (b) Adaptive Spatial Attention Module (ASM). Given an input feature tensor $F \in \mathbb{R}^{H\times W\times C}$, the ACM first applies global average pooling (GAP) and global max pooling (GMP) to generate channel descriptors, which are passed through a shared multi-layer perceptron (MLP) with ReLU activation and combined to produce channel attention weights $M_C(F) \in \mathbb{R}^{1\times 1\times C}$ via a sigmoid function. In parallel, the ASM computes spatial attention by aggregating channel-wise average pooling and max pooling features, concatenating them, and processing the result through a depth-wise separable convolution followed by a sigmoid activation to obtain spatial attention weights $M_S(F) \in \mathbb{R}^{H\times W\times 1}$. The final refined feature representation is obtained through sequential element-wise re-weighting of the input feature map using both channel and spatial attention maps, ensuring a complete closed-loop transformation from input features to enhanced output features.

classification performance. The AFN adopts a CNN model that incorporates SAAN, which employs ACM and ASM to get both global and local contextual information across channels and spatial patterns associated with fire. Leveraging depth-wise separable convolution with SAAN provides more effective parameterization and enhances discriminative power, making it effective for detecting fires. Unlike conventional detection-based frameworks that focus on localizing fire regions using bounding boxes, the proposed model addresses fire scene classification, where the objective is to assign a global semantic label to the entire image. This formulation aligns with UAV-based monitoring scenarios where rapid scene-level decision-making is required. **Figure 1** represents the entire framework flow of the proposed model.

3.1 Architecture Overview

Early and timely fire detection is important for saving human lives and their property. For this purpose, various ML and DL-based models were deployed. However, these approaches have some limitations as discussed in the introduction section, some of the restrictions are high False Positive Rates, low performance, and feature engineering. Therefore, this research paper introduces an AFN architecture composed of a deep intelligent model followed by SAAN that is further composed of two sub-modules, ACM and ASM. The ASM focuses on relevant spatial features, such as shape, color, patterns, and other key indicators of fires. While ACM is used to capture informative features across frames. This means that the system can detect changes in the fire's size, shape, and location over time, as well as distinguish between different types of fires (e.g. wildfires, building fires, etc). All these techniques work together to develop an accurate fire detection and classification system that can assist in protecting urban property and the environment from the devastating effects of fires. The technical details of the proposed framework are explored in the coming section. The SAAN architecture is presented in **Figure 2** with a diverse collection of feature fusion that enhances the model discriminative power for fire scenes.

3.2 Backbone Architecture Model

Deep learning-based visual recognition models, including EfficientNet, ResNet, and Xception, have been widely adopted across different computer vision tasks. Pre-trained models are particularly useful because they provide transferable visual representations that can be fine-tuned for domain-specific tasks with limited task-specific data. In fire-scene classification, this property is important because fire patterns vary substantially with illumination, background, smoke density, camera viewpoint, flame scale, and environmental conditions. In this study, ConvNeXt-B was selected as the backbone architecture of the proposed Adaptive Fire Network (AFN). ConvNeXt modernizes conventional convolutional network design by incorporating architectural principles inspired by recent high-performing vision models while retaining the computational advantages of convolutional operations[37]. The ConvNeXt family includes multiple variants, such as ConvNeXt-T, ConvNeXt-S, ConvNeXt-B, ConvNeXt-L, and ConvNeXt-XL, which differ in depth and computational capacity. Among these variants, ConvNeXt-B provides an effective balance between feature-learning capability and computational cost. The ConvNeXt-B backbone receives input images resized to $224 \times 224 \times 3$ and extracts hierarchical feature representations through convolutional blocks. Lower-level layers capture visual cues such as edges, color transitions, and flame-like textures, whereas deeper layers encode higher-level semantic information related to fire regions, smoke patterns, and contextual background structures. These features are then refined by the proposed Scene-Aware Attention Network (SAAN), which consists of the Adaptive Channel Module (ACM) and Adaptive Spatial Module (ASM). Through this integration, AFN emphasizes fire-relevant channel responses and spatial regions while reducing the influence of visually similar non-fire patterns. The model was trained for binary fire/non-fire classification using binary cross-entropy as the loss function and the Adam optimizer. Training was conducted for 32 epochs with a batch size of 16 and an initial learning rate of 1×10^{-4}. The network parameters were updated through backpropagation to minimize the classification loss. Since binary cross-entropy and Adam are standard components in deep learning-based classification models, their mathematical derivations are omitted to maintain the focus on the proposed AFN architecture.

3.3 Scene-Aware Attention Network

The Scene-Aware Attention Network (SAAN) is a key component of the proposed Adaptive Fire Network (AFN). It is designed to refine the feature representations extracted by the ConvNeXt-B backbone by emphasizing informative channel responses and spatially relevant fire regions. This refinement is particularly important for UAV-assisted and drone-relevant fire monitoring scenarios, where fire regions may appear at different scales, under varying viewpoints, and against complex backgrounds such as forests, urban areas, roads, buildings, smoke, sunlight, and fire-like objects. SAAN consists of two sequential modules: the Adaptive Channel Module (ACM) and the Adaptive Spatial Module (ASM). The ACM models inter-channel dependencies to identify discriminative feature channels, whereas the ASM models spatial saliency to highlight fire-relevant regions in the feature map.

Let $\theta \in \mathbb{R}^{H \times W \times C}$ denote the input feature map to the ACM, where H, W, and C represent the height, width, and number of channels, respectively. To obtain channel descriptors, global average pooling and global max pooling are independently applied along the spatial dimensions. These operations generate two complementary channel descriptors, as follows:

$$\theta_{\text{CA}}^{\text{avg}} = \text{GAP}(\theta) \tag{1}$$

$$\theta_{\text{CA}}^{\text{max}} = \text{GMP}(\theta) \tag{2}$$

where $\text{GAP}(\cdot)$ and $\text{GMP}(\cdot)$ denote global average pooling and global max pooling, respectively. The average-pooled descriptor captures the overall channel-wise response, whereas the max-pooled descriptor preserves the most activated channel information. These descriptors are useful in UAV-assisted fire monitoring because fire-related patterns may occupy only a small portion of the image, particularly when the camera observes the scene from a distant or elevated viewpoint.

The two channel descriptors are then passed through shared fully connected layers to generate channel-attention responses. Specifically, each descriptor is processed by the first fully connected layer, followed by a rectified linear unit (ReLU) activation function, and then passed through the second fully connected layer. The max-pooling and average-pooling branches are formulated as:

$$M_{\text{max}} = fc^2 \left(\Theta \left(fc^1 \left(\theta_{\text{CA}}^{\text{max}}\right)\right)\right) \tag{3}$$

$$M_{\text{avg}} = fc^2 \left(\Theta \left(fc^1 \left(\theta_{\text{CA}}^{\text{avg}}\right)\right)\right) \tag{4}$$

where fc^1 and fc^2 denote the first and second fully connected layers, respectively, and Θ denotes the ReLU activation function. The two channel-attention responses are then fused and normalized using a sigmoid activation function:

$$M_{\text{CA}}(\theta) = \sigma \left(M_{\text{max}} \oplus M_{\text{avg}}\right) \tag{5}$$

where σ denotes the sigmoid activation function and $\oplus$ denotes element-wise addition between the two channel-attention responses. The resulting channel-attention map $M_{\mathrm{CA}}(\theta)$ is applied to the input feature map through element-wise multiplication to obtain the channel-refined feature map:

$$F_{\mathrm{CA}} = M_{\mathrm{CA}}(\theta) \odot \theta \tag{6}$$

where $\odot$ denotes element-wise multiplication. The channel-attention map is broadcast along the spatial dimensions so that informative channels are emphasized while less relevant channel responses are suppressed. This formulation clearly separates the max-pooling and average-pooling branches and ensures that each branch uses its corresponding descriptor.

The ASM further refines the channel-enhanced feature map F_{CA} by estimating spatial attention. While ACM focuses on identifying important channels, ASM focuses on locating informative spatial regions associated with fire patterns. This is important in UAV-assisted fire monitoring because fire regions can be partially occluded by smoke, appear at small spatial scales, or be surrounded by visually similar background patterns. Given F_{CA}, average pooling and max pooling are applied along the channel dimension to generate two spatial descriptors:

$$F_{\mathrm{S}}^{\mathrm{avg}} = \mathrm{AvgPool}_c(F_{\mathrm{CA}}) \tag{7}$$

$$F_{\mathrm{S}}^{\mathrm{max}} = \mathrm{MaxPool}_c(F_{\mathrm{CA}}) \tag{8}$$

where $\mathrm{AvgPool}_c(\cdot)$ and $\mathrm{MaxPool}_c(\cdot)$ denote average pooling and max pooling along the channel dimension, respectively. These two spatial descriptors are concatenated and processed by convolutional layers to generate the spatial-attention map:

$$M_{\mathrm{SA}}(F_{\mathrm{CA}}) = \sigma\left(f^{1\times 1}\left(\Theta\left(f^{3\times 3}\left(\Theta\left(f^{7\times 7}\left(\left[F_{\mathrm{S}}^{\mathrm{avg}}; F_{\mathrm{S}}^{\mathrm{max}}\right]\right)\right)\right)\right)\right)\right) \tag{9}$$

where [;] denotes concatenation, and $f^{7\times 7}$, $f^{3\times 3}$, and $f^{1\times 1}$ denote convolutional operations with kernel sizes of 7×7, 3×3, and 1×1, respectively. The sigmoid function normalizes the spatial-attention map to the range $[0,1]$. The final SAAN-refined feature map is obtained by applying the spatial-attention map to the channel-refined feature map:

$$F_{\mathrm{SAAN}} = M_{\mathrm{SA}}(F_{\mathrm{CA}}) \odot F_{\mathrm{CA}} \tag{10}$$

Thus, SAAN sequentially performs channel-wise and spatial-wise feature refinement. The ACM emphasizes discriminative channels related to fire-specific visual patterns, while the ASM highlights spatial regions that are more relevant for distinguishing fire from visually similar non-fire scenes. This dual refinement is beneficial for UAV-assisted and drone-relevant fire monitoring because it helps the model remain attentive to fire regions despite scale variation, viewpoint changes, background clutter, and fire-like distractors. The final refined representation F_{SAAN} is then forwarded to the subsequent classification stage of AFN.

3.4 Adaptive Fire Network

The Adaptive Fire Network (AFN) integrates the ConvNeXt-B backbone with the proposed Scene-Aware Attention Network (SAAN) to perform binary fire/non-fire classification. The overall architecture follows a sequential design in which the input image is first processed by ConvNeXt-B to extract hierarchical visual features. These backbone features are then refined by SAAN, which applies channel-wise and spatial-wise attention through the Adaptive Channel Module (ACM) and the Adaptive Spatial Module (ASM), respectively. The refined feature representation is subsequently forwarded to the classification head to generate the final prediction.

The purpose of AFN is to combine the strong feature extraction capability of ConvNeXt-B with the adaptive refinement ability of SAAN. ConvNeXt-B captures both low-level and high-level visual patterns, while SAAN emphasizes fire-relevant channel responses and spatial regions. This design is useful for UAV-assisted and drone-relevant fire monitoring scenarios, where fire regions may appear at different scales, under varying viewpoints, and against complex backgrounds such as smoke, sunlight, vegetation, buildings, and fire-like objects.

In the final architecture, the model is initialized with pre-trained ConvNeXt-B weights and fine-tuned using fire-scene images. Task-specific classification layers are added after the SAAN-refined feature representation to adapt the model to binary fire classification. This integrated design allows AFN to preserve the general visual representation ability of the backbone while improving its sensitivity to fire-specific patterns through attention-based feature refinement.

4 Experimental Results

In this section, a comprehensive description of the experimental setup is provided to evaluate the performance of the proposed model. It includes details regarding the experimental setup, datasets, training procedure, and evaluation metrics. Furthermore, quantitative and qualitative analyses are also conducted where the outcome of the proposed model is fairly compared with SOTA methods to demonstrate its effectiveness. Additionally, ablation studies are performed to assess the contribution of individual components of the proposed model to its overall performance.

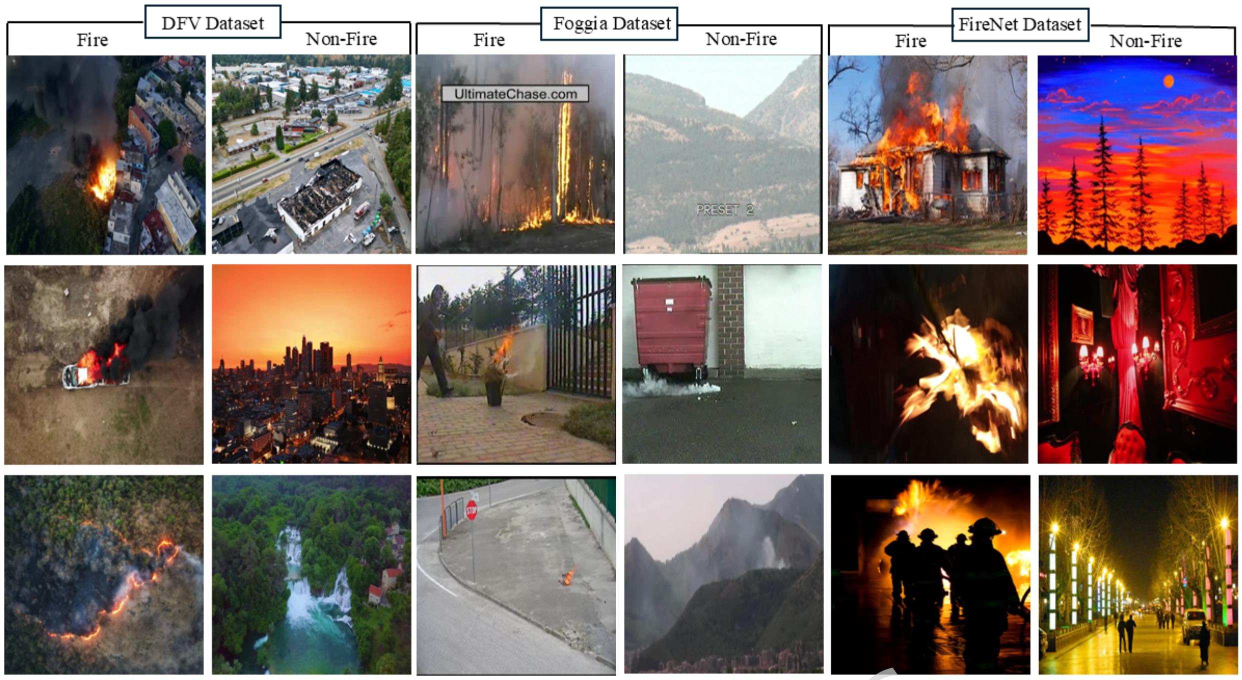

Figure 3. Representative samples from the DFV, Foggia, and FireNet datasets. The DFV dataset consists of UAV-captured drone-view fire and non-fire images, while the Foggia and FireNet datasets are heterogeneous benchmark datasets containing fire and non-fire scenes from surveillance, ground-level, and aerial-like perspectives. This figure highlights the diversity in viewpoint, illumination, background complexity, and fire scale across datasets, supporting robust evaluation of model generalization in both UAV-specific and cross-domain fire classification scenarios.

4.1 Datasets and Evaluation Metrics

The AFN is evaluated and compared with SOTA methods using standard evolution metrics, including accuracy, recall, F1-score, precision, False Positive Rate, and false negative rate, which are widely used in the literature to assess the fire detection model performance. After a comprehensive evaluation, we concluded that AFN could detect fire accurately and precisely. Fire is variable in nature and can change with environmental conditions, types, and intensity. Therefore, developing a reliable and adaptable fire detection model requires a dataset with a large collection of fire scenarios that covers heterogeneous environmental conditions and fire patterns. For this purpose, our research utilizes three datasets for experimental and evaluation purposes, allowing the model to be trained and validated against an extensive and diverse sample of actual fire situations. This approach enhances the model's reliability and generalizability in real-world applications. **Figure** 3 represents the sample images source from the Drone Fire Vision , Foggia's[38] and FireNet[39] datasets, while the statistical information about data is reported in **Table** 1.

- **Drone Fire Vision:** The proposed Drone Fire Vision (DFV) dataset was constructed from a diverse collection of aerial and drone-related visual data sources, including publicly available videos and images retrieved from platforms such as YouTube, Facebook, and Google. To enhance data diversity, additional non-fire samples were incorporated from the VisDrone dataset[40]. The collected videos were decomposed into frames using a structured preprocessing pipeline, where redundant and low-quality frames were removed to ensure consistency, visual clarity, and relevance to fire classification tasks. A representative subset of informative frames was then selected to form the final dataset used in this study. The DFV dataset consists of 6000 images, including 4200 training and 1800 testing samples, with an equal balance between fire and non-fire categories. To further characterize dataset diversity, we performed a quantitative statistical analysis of key scene attributes. The results indicate that the dataset is primarily composed of daytime imagery (4927 images, 82.1%), while 1073 images (17.9%) correspond to nighttime conditions. In terms of background context, 745 fire-related samples (24.8%) are associated with vegetation or forest environments, whereas 2255 samples (75.2%) represent urban or mixed scenes. Additionally, fire instances exhibit significant scale variation, with 200 images (3.3%) corresponding to small-scale fires, 1212 images (20.2%) to medium-scale fires, and 4588 images (76.5%) to large-scale fire events. These

Table 1. The detailed statistics of training, validation and test images of Foggia's, DFV, and FireNet datasets

Dataset	Dataset	Train	Valid	Test
Foggia's	Fire	5266	436	1316
	Non-Fire	5302	428	1288
	Total	10568	864	2604
DFV	Fire	2100	-	900
	Non-Fire	2100	-	900
	Total	4200	-	1800
FireNet	Fire	1124	-	593
	Non-Fire	1301	-	278
	Total	2425	-	871

distributions demonstrate that DFV contains substantial variability in illumination, environmental context, and fire scale, supporting its suitability for robust and diverse fire classification under UAV-based scenarios.

- **Foggia's:** It was created by[38], known as Foggia's Dataset, focuses on fire detection using videos. This dataset contains 31 videos, 14 of which consist of fire scenes while the remaining 17 do not. Utilizing the frame extraction method described by[36], a total of 14,036 frames were extracted from these videos. These frames were then categorized into two classes: fire and non-fire. Each class consists of an equal number of frames, 7,018, ensuring a balanced dataset for training and evaluating fire detection algorithms. Importantly, the dataset primarily contains surveillance-style and ground-level visual perspectives rather than UAV-specific aerial imagery. This characteristic makes it particularly useful for evaluating the robustness of fire detection models under conventional camera-based monitoring conditions and contributes to assessing cross-domain generalization capability when combined with UAV-oriented datasets.

- **FireNet:** The FireNet dataset is a challenging benchmark dataset for fire and non-fire classification tasks, introduced by Jadon et al[39]. It contains a diverse collection of images gathered from multiple sources, including publicly available web repositories and surveillance footage. The dataset consists of 1,124 fire images and 1,301 non-fire images, ensuring moderate class balance. Additionally, a separate test set of 871 images is provided, including 593 fire and 278 non-fire samples. The dataset includes a wide range of visual scenarios such as indoor fire incidents, outdoor urban environments, and complex lighting conditions where fire-like visual patterns may appear. Unlike UAV-specific datasets, FireNet includes heterogeneous and mixed-view imagery, making it a strong benchmark for evaluating model robustness under diverse real-world conditions and for testing generalization beyond controlled aerial perspectives.

The DFV dataset is specifically designed for UAV-based fire monitoring scenarios and consists of drone-view imagery. In contrast, the Foggia and FireNet datasets serve as widely used benchmark datasets and include heterogeneous visual perspectives, ranging from ground-level surveillance to aerial-like views. This multi-dataset setup enables a comprehensive evaluation of model generalization across different visual domains and ensures robust performance assessment under both UAV-specific and general fire detection scenarios.

4.2 System Configuration and Implementation Detail

In this research work Python version 3.8, Keras DL framework with TensorFlow backend, Intel® Core™ i9 14900K CPU clocked at 5.0 GHz paired with an NVIDIA GeForce RTX 3090 GPU were used. The GPU possesses exceptional deep-learning acceleration abilities and is equipped with a substantial 24GB of onboard memory. Its high floating point arithmetic performance enables it to achieve 35.58 tera-floating-point operations, ensuring efficient high-performance computing for demanding Artificial Intelligence tasks. 32 epochs were used for training of proposed and ablation models with inputs size of 224×224×3, 16 batch size, optimizer Adam and learning rate of 1e-4. These configuration parameters were meticulously carefully chosen based on extensive experimentation to optimize model performance

4.3 Ablation Study

In this experimental segment, two types of ablation studies are performed to evaluate the most effective configuration of the AFN network. These investigations involved experimentation with various pre-trained CNN models to select the best architecture for the backbone model applying the FireNet dataset[39]. Furthermore, for fair evaluation, we have also conducted experiments over various integration of attention modules and assessed the effectiveness of the AFN across different compositions of ACM and ASM modules on the Drone dataset. The findings of the backbone model are reported in **Table 2** while the results of the Adaptive attention module are given in **Table 3** and elaborated in the subsequent sections.

Table 2. An ablation study was conducted over several backbone models on the FireNet dataset in terms of Accuracy (ACC), F1-score (F1), Recall (R), Precision (P), False Positive Rate (FPR) and False Negative Rate (FNR).

Method	ACC	F1	R	P	FPR	FNR
EfficientNetV2B3	85.99	89.22	93.69	85.16	26.51	6.31
EfficientNetV2S	86.12	90.64	94.18	87.35	33.94	5.82
ConvNeXtTiny	90.93	92.94	98.86	87.69	21.16	1.14
EfficientNetV2B2	92.42	94.33	96.15	92.58	14.67	3.85
EfficientNetV2B1	92.88	94.56	98.54	90.89	16.67	1.46
ConvNeXtSmall	93.23	94.81	99.08	90.89	16.51	0.92
ConvNeXtBase	94.14	95.51	99.82	91.57	15.29	0.18

Table 3. An ablation study of the proposed model on the DFV dataset in terms of accuracy (ACC), F1-score (F1), recall (R), and precision (P). The red color indicates the best architecture.

Method	ACC	F1	R	P
BM	82.06	81.80	82.97	80.67
BM + ACM	83.39	85.04	77.34	94.44
BM + ASM	83.78	85.13	78.57	92.89
BM + ACM-I + ASM	84.06	85.44	78.62	93.56
BM + ACM-II + ASM	88.39	89.24	83.13	96.33
AFN	**90.78**	**91.27**	**86.63**	**95.99**

4.3.1 Investigating the Backbone Architecture

The first ablation study was conducted over various pre-trained CNN models to select the best backbone architecture for the AFN network. The compared model includes EfficientNetV2B3, EfficientNetV2S, ConvNeXtTiny, EfficientNetV2B2, EfficientNetV2B1, ConvNeXtSmall and ConvNeXtBase, where every model results are given in **Table 2** over the FireNet dataset[39]. The EfficientNetV2B3 performed very vaguely and achieved the lowest accuracy 85.99% among the other pre-trained CNN models. In contrast, the ConvNeXtBase surpasses the existing pre-trained models by accomplishing the highest accuracy of 94.14%, 95.51% F1-Score, 99.82% Recall, 91.57% precision and lowest FNR while the false positive alarm rate of 15.2% which is lowest from remaining pre-trained CNN model but slightly high then EfficientNetV2B2 which is 0.62% but the accuracy and other parameters of EfficientNetV2B2 are not acceptable. The improved performance of the ConvNeXtBase indicates its effectiveness for the effective configuration of AFN. This ablation study summarized that pre-trained CNN Networks have restricted and limited performance when utilizing the fire dataset. Nonetheless, the most improved performance was achieved by implementing the AFN leveraging the DFAN module in terms of accuracy and minimizing the False Positive Rate and false negative rate. In **Table 2**, the red color signifies the best performance while the light blue font color shows the second best and the dark blue text color specifies the lowest performance of the pre-trained model on the FireNet dataset[39].

Table 4. Performance comparison of conventional and proposed attention mechanisms incorporated into the ConvNeXtBase baseline on the FireNet dataset.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
BM	94.14	91.57	99.82	95.51
BM + SE	95.52	96.01	97.47	96.74
BM + ECA	96.21	96.51	97.98	97.24
BM + CBAM	96.79	97.01	98.32	97.66
BM + ACM	97.24	97.50	98.49	98.00
BM + ASM	97.70	97.84	98.83	98.33
BM + ACM-I + ASM	98.05	98.17	99.00	98.58
BM + ASM-II + ASM	98.40	98.50	99.16	98.83
AFN (Proposed)	**98.97**	**99.16**	**99.32**	**99.24**

.

Table 5. Performance comparison of the proposed model with SOTA methods over Foggia's dataset. The best score is represented in blue with bold.

	Methods	Foggia's Dataset [38]		
		FPR	FNR	ACR
Traditional Models	FSD(RGB)[41]	41.18	7.14	74.20
	FD-GCM[19]	29.41	0	83.87
	FSD(YUV)[41]	17.65	7.14	87.10
	EFD-IP[11]	11.76	14.29	87.10
	CM-FDM[7]	5.88	14.29	90.32
	FD-CV[42]	13.33	0	92.86
	FD-CSM[38]	11.67	0	93.55
Deep Learning	ANetFire[43]	9.07	2.13	94.39
	GNetFire[44]	0.054	1.50	94.43
	CNNFire[45]	8.87	2.12	94.50
	EMNFire[21]	0	0.14	95.86
	ICA-K[46]	4.83	4.53	95.32
	LCHP-DLM[47]	0	0.80	99.19
	ViT-B/32[48]	2.15	1.02	94.03
	CNN-A CHOR[49]	8.65	1.92	94.12
	FireNet[39]	1.23	2.25	96.53
	SE-EFFNet[50]	0.04	0.03	97.2
	STN-CNN[51]	3.68	2.46	96.2
	CANet[52]	0	0.02	98.34
	DFAN[36]	0	0.58	99.60
	FasterNetFire[53]	0.12	0	99.66
	CA-Swin[54]	0.14	0	99.87
	AFN	**0.008**	**0**	**99.96**

4.3.2 Impact of Integrated Adaptive Attention Module

The second ablation study was conducted to choose the best Adaptive attention module for the AFN network. This study also examines the impact of each module, more particularly, each attention module is combined with the ConvNeXtBase, followed by training and evaluation. The empirical study began with a systematic analysis of a basic backbone model and progressively incorporated individual ACM with backbone model utilizing Drone Dataset accomplishing the accuracy of 82.06% and 83.39% respectively, then we expanded our investigation and evaluated the performance of ASM with backbone model and get the performance of 83.78% that indicates that simple pre-trained model and individual ACM and ASM is not enough for completed fire detection and classification. Next, we implemented an individual ACM along with an ASM plus backbone model leveraging depth-wise separable convolution, achieving an accuracy of 84.06% Subsequently, integrating a second ACM with the ASM and ConvNeXt-B led to an improved performance of 88.39%. By integrating the various Adaptive attention model the performance of the model is enhanced by an accuracy of 6.33% from the backbone model, 5% from individual ACM integrated with the backbone model, and 4.61% increase from ASM integrated with the backbone model, but still, it is not acceptable for fire detection and classification. To get a robust and efficient model, we fused both dual ACM and the ASM with the ConvNeXt-B architecture using depth-wise separable convolution, resulting in a strong performance of 90.78% for fire classification and localization using the drone dataset. In summary, AFN achieved an accuracy of 8.72% from the pre-trained Backbone model, 7.39% accuracy higher than BACM, a 7.00% accuracy increase from BASM, BACSMI achieved a low performance of 6.72% and lastly BACSMI gets 2.39% lower accuracy form AFN network. Our findings, as outlined in **Table 3**, demonstrate that backbone models augmented with dual ACM, ASM using depth-wise separable convolutional surpass methodologies relying solely on deep features. This advancement maintains particular significance, given the inherent challenges of fire classification compared to standard ImageNet tasks. The integration of dual Adaptive attention modules improves the extraction of distinctive object features, thereby elevating the accuracy of fire scene classification. Among the baseline models, ConvNeXt-Base features combined with a SoftMax classifier exhibited improved performance owing to their exceptional feature extraction capabilities. Moreover, as shown in **Table 3**, our proposed model consistently outperforms others, while the simple backbone model and other Adaptive attention module integrations exhibit inferior performance due to their limited adaptability and applicability of features to fire scene classification.

Table 6. The proposed modelś performance is compared with existing SOTA approaches using FireNet, where the best and second best accuracy is denoted by blue and red color.

Method	ACC	F1	R	P	FPR	FNR
FireNet[39]	93.91	95	94	97	1.95	4.13
Nguyen[55]	96.08	-	-	-	-	-
UFS-Net[56]	97.13	-	-	98.29	2.56	3.5
DWConv[57]	97.14	-	-	-	-	-
AFN	**98.97**	**99.24**	**99.32**	**99.16**	**1.79**	**0.68**

To further evaluate the effectiveness of the proposed SAAN module, an additional horizontal comparison was conducted on the FireNet dataset by incorporating representative attention mechanisms into the ConvNeXtBase backbone. Specifically, SE, ECA, and CBAM were selected as widely used attention baselines and evaluated under the same experimental settings to ensure a fair comparison. As shown in **Table 4**, the original ConvNeXtBase backbone achieved an accuracy of 94.14%, precision of 91.57%, recall of 99.82%, and F1-score of 95.51%. With the addition of SE, the ConvNeXtBase model achieved 95.52% accuracy and 96.74% F1-score. The ECA-based variant further improved the accuracy and F1-score to 96.21% and 97.24%, respectively. Among the conventional attention baselines, the CBAM-enhanced ConvNeXtBase model obtained the strongest performance, reaching 96.79% accuracy, 97.01% precision, 98.32% recall, and 97.66% F1-score. The proposed attention components yielded further improvements over these mainstream attention mechanisms. The ACM-based ConvNeXtBase model achieved 97.24% accuracy and 98.00% F1-score, whereas the ASM-based configuration obtained 97.70% accuracy and 98.33% F1-score. When ACM-I was combined with ASM, the performance increased to 98.05% accuracy and 98.58% F1-score. The extended configuration using ACM-II and ASM further raised the accuracy and F1-score to 98.40% and 98.83%, respectively. The complete AFN model achieved the best overall performance, with 98.97% accuracy, 99.16% precision, 99.32% recall, and 99.24% F1-score. Compared with the original ConvNeXtBase backbone, AFN improved accuracy by 4.83 percentage points and F1-score by 3.73 percentage points. Compared with the CBAM-enhanced ConvNeXtBase model, which was the strongest conventional attention baseline in this comparison, AFN improved accuracy by 2.18 percentage points and F1-score by 1.58 percentage points. These results demonstrate that the proposed SAAN module provides more effective feature refinement than directly adopting existing generic attention mechanisms. The progressive improvements achieved by the ACM-based, ASM-based, combined ACM–ASM, and complete AFN configurations further indicate that channel-wise and spatial feature refinement provide complementary benefits for fire-scene classification.

5 Performance Comparison with the SOTA Approach

In this section, detailed experiments were performed over the three datasets including Foggia's, Drone Fire Vision and FireNet to elucidate the proposed AFN Network performance. These evaluations of the AFN network are categorized into two types, quantitative and quality analysis with SOTA.

5.1 Quantitative Results Analysis With SOTA

The comparative analysis presented sheds light on the efficacy and significance of various methodologies for fire detection, encompassing both TML and DL approaches, particularly concerning Foggia's and FireNet datasets. Through the lens of critical performance metrics such as FPR, FNR, and ACR, the performance of these methodologies is thoroughly evaluated. Traditional models, including FSD(RGB)[41], FD-GCM[19], FSD(YUV)[41], EFD-IP[11], CM-FDM[58], FD-CV[42], and FD-CSM[38], demonstrate commendable performance across different metrics using Foggia's dataset. However, each model exhibits specific strengths and weaknesses. For instance, while FD-GCM excels in minimizing missed detections with a remarkable FNR of 0%, it concurrently struggles with a relatively higher false alarm rate, as evidenced by its FPR of 29.41%. Conversely, CM-FDM showcases promising capabilities in reducing false alarms, boasting a notably low FPR of 5.88%. As a whole the results of FSD (RGB)[41] worst in terms of accuracy which is 74.20%, in contrast, the FD-CSM performed well as compared to other methods and achieved an accuracy of 93.55% indicating that it enhanced the performance of the model up to 19.35% as compared to the FSD(RGB)[41].

In contrast, DL methodologies herald a new era of precision and robustness in fire detection. Models such as ANetFire[43], GNetFire[44], CNNFire[45], EMNFire[21], ICA-K[46], LCHP-DLM[45], ViT-B/32[48], CNN-A CHOR[49], and DFAN[36] consistently outperform traditional models across FPR and FNR metrics utilizing the Foggia dataset. ANetFire[43], for instance, stands out with an exceptional FPR of 9.07%, an equally minimal FNR of 2.13% and achieved higher accuracy as compared to the TML approaches, highlighting its effectiveness in minimizing false alarms while ensuring accurate fire detection, but still its results are not acceptable for AUV based fire detection and implementation. Due to the importance of the fire domain,

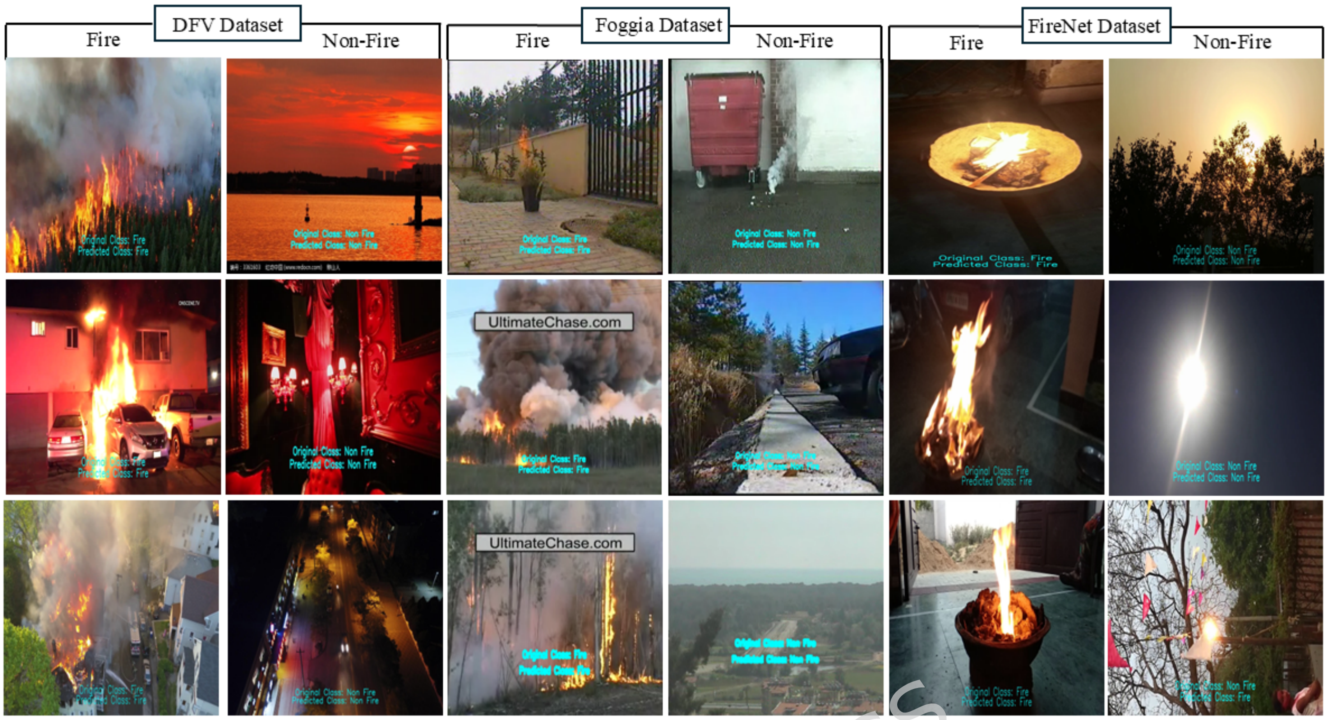

Figure 4. Predicted qualitative results of the AFN using DFV, Foggia and FireNet datasets.

other researchers have proposed various models for correct and precise fire detection. EMNFire[21] and LCHP-DLM[47] develop a model whose minimized FRP rate is 0 but still the other parameter needs improvement. Finally, DFAN[36] get the highest accuracy of 99.60% as compared to the previous methods which is 5.48% greater than CNN-A CHOR[49]. More recent methods, including FasterNetFire[53] and CA-Swin[54], also report very high accuracy rates of 99.66% and 99.87%, respectively, with a false negative rate of 0%. Our proposed AFN network presents a strong performance in terms of accuracy which is 99.96%, with the lowest FNR of 0 among all the SOTA and FPR is 0.008% which is the lowest in all the SOTA techniques but slightly greater than EMNFire[21], LCHP-DLM[47] and DFAN[36] as shown in **Table 5**.

In the domain of fire detection, FireNet is considered one of the most prominent and challenging datasets among others. We considered the same setup for the experiments as mentioned in[39] dataset. **Table 6** provides a comparative analysis of several methods based on their performance metrics, including accuracy, F1 score, precision, recall, FPR, and FNR. FireNet[39] achieves an accuracy of 93.91%, with a balanced F1 score of 95%, demonstrating high precision (97%) and recall (94%), albeit with a higher FNR of 4.13%. Its performance is good, but we need to improve the performance by reducing the FPR and achieving high accuracy. DWConv[56] and Nguyen et al.[55] methods reported accuracies of 97.14% and 96.08%, respectively. Nguyen[55] increased by 2.17% as compared to FireNet[39] while DWConv[57] achieved 3.23% higher accuracy as compared to FireNet[39] and 1.06% Nguyen[55]. UFS-Net[56] achieves an accuracy of 97.13% with relatively high precision (98.29%) and a moderate FPR of 2.56%. However, the FNR is still high at 3.5%. its accuracy is good but still, we need to reduce the FPR and FNR. In contrast, the proposed AFN Network outperforms other methods with an accuracy of 98.97% and a notably high F1 score of 99.24%, indicating excellent precision of 99.16% and recall of 99.32%. Moreover, the AFN Network demonstrates low false positive 1.79% and false negative 0.68% rates, highlighting its reliability and effectiveness in fire detection and classification tasks. To summarize, the performance of our proposed model AFN increased the accuracy by 5.06% as compared to FireNet[39], 2.89% from Nguyen[55], 1.84% from UFS-Net[56] and 1.83% from DWConv[57]. Overall, AFN achieved the highest performance among the compared methods, offering improved performance across various evaluation metrics. This high precision not only underscores the improved performance of the proposed model but also underscores its potential to contribute to fire detection systems, promising enhanced public safety and property preservation.

5.2 Qualitative Results Analysis

The proposed model is also evaluated by quantitative analysis and investigated its performance is investigated by distinguishing the image labeled as fire or non-fire. The image samples are passed from all three datasets and the proposed model accurately predicted the image containing fire and labeled it as Fire while the image that does not contain the fire is labeled as non-fire as

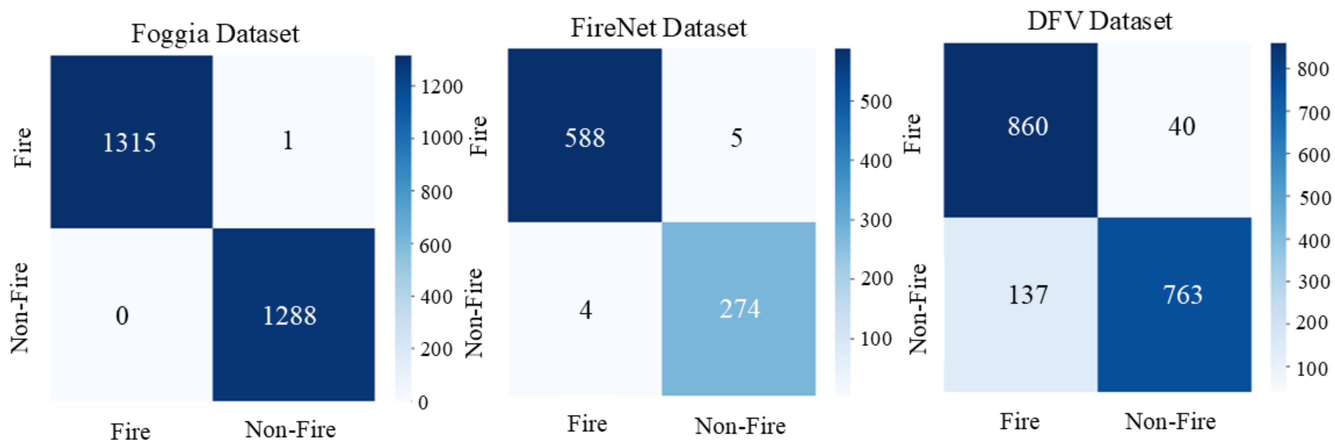

Figure 5. The performance of the AFN over tested data using Foggia, FireNet, and DFV datasets.

demonstrated in **Figure 4**. The image set of the first two columns is collected from the drone dataset and the proposed model has accurately identified as fire and non-fire and labeled according to the ground truth label. The third and fourth column represents the image of Foggia's dataset which correctly recognized fire images as fire while non-fire images as non-fire. The fourth and fifth columns are selected from the FireNet dataset, and the AFN network is appropriately detected and given labels according to its original label. The first and second images of a non-fire column in the drone dataset completely look like fire but the proposed model has accurately recognized it as non-fire, similarly, the first image of the fire column in Foggia's dataset looks like non-fire because there is very little region of fire in the image and It is very difficult for ordinary fire detection model to detect small fire b/c its shape change continuously but the AFN has detected fire in the image and label as fire which show its robustness of the AFN. In the FireNet dataset, the first and third images of non-fire look like fire but AFN correctly detects and categorizes it into non-fire. This functionality is obtained by the operating mechanism of dual ACM and ASM modules leveraging the dept-wise separable convolution in the feature extraction process. **Figure 4** demonstrates that the proposed model is capable in challenging conditions and it can accurately and precisely predict the fire in the images.

5.2.1 Confusion Matrix

The confusion matrix performs as a core tool in evaluating the performance of detection and classification models, providing an in-depth breakdown of predicted labels versus actual class labels. Its significance lies in its ability to quantify the model's effectiveness in distinguishing between different classes, thereby guiding further analysis and model refinement. Diagonal elements signify correct predictions, while off-diagonal elements denote incorrect predictions. Higher values along the diagonal indicate superior model performance with fewer miscategorized. The results of our proposed model showcase impressive performance metrics, boasting high accuracy along with the lowest false positive and negative rates in all three datasets as illustrated in **Figure 5**. Its achievement underscores the efficacy of our approach in accurately distinguishing between fire and non-fire instances. The incorporation of dual ACM coupled with ASM using depth-wise separable convolution has played a pivotal role in achieving these outcomes. By leveraging these advanced techniques, our model effectively focuses on relevant features while mitigating noise and irrelevant information, thereby enhancing its discriminative power. This combination empowers our model to achieve competitive performance in fire detection tasks, setting a promising precedent for its practical application in real-world scenarios.

To select the optimal backbone model for the AFN network we have performed a wide range of ablation studies on pre-trained mode utilizing the FireNet dataset. FireNet is considered the most challenging dataset, and the confusion matrix plays a important role in the analysis of a model. Among the pre-trained models tested, ConvNextBase demonstrated improved performance, exhibiting minimized FP and FN results as compared to the other pre-trained CNN networks. Conversely, EfficientNetV2B3 displayed comparatively poorer performance in terms of accuracy, high False Positive Rate and false negative rate. The efficacy of ConvNextBase can be attributed to its robust architecture, which effectively discriminates between fire and non-fire instances, thus reducing false alarms. Conversely, the suboptimal performance of EfficientNetV2B3 may stem from its architecture's limitations in capturing intricate fire-related features, leading to increased misclassifications. These findings underscore the significance of architectural design in enhancing fire detection accuracy. **Figure 6** shows the visual representation of the ablation study conducted on the FireNet dataset using a pre-trained CNN model. **Figure 7** illustrates the visual representation of three datasets using heat map techniques.

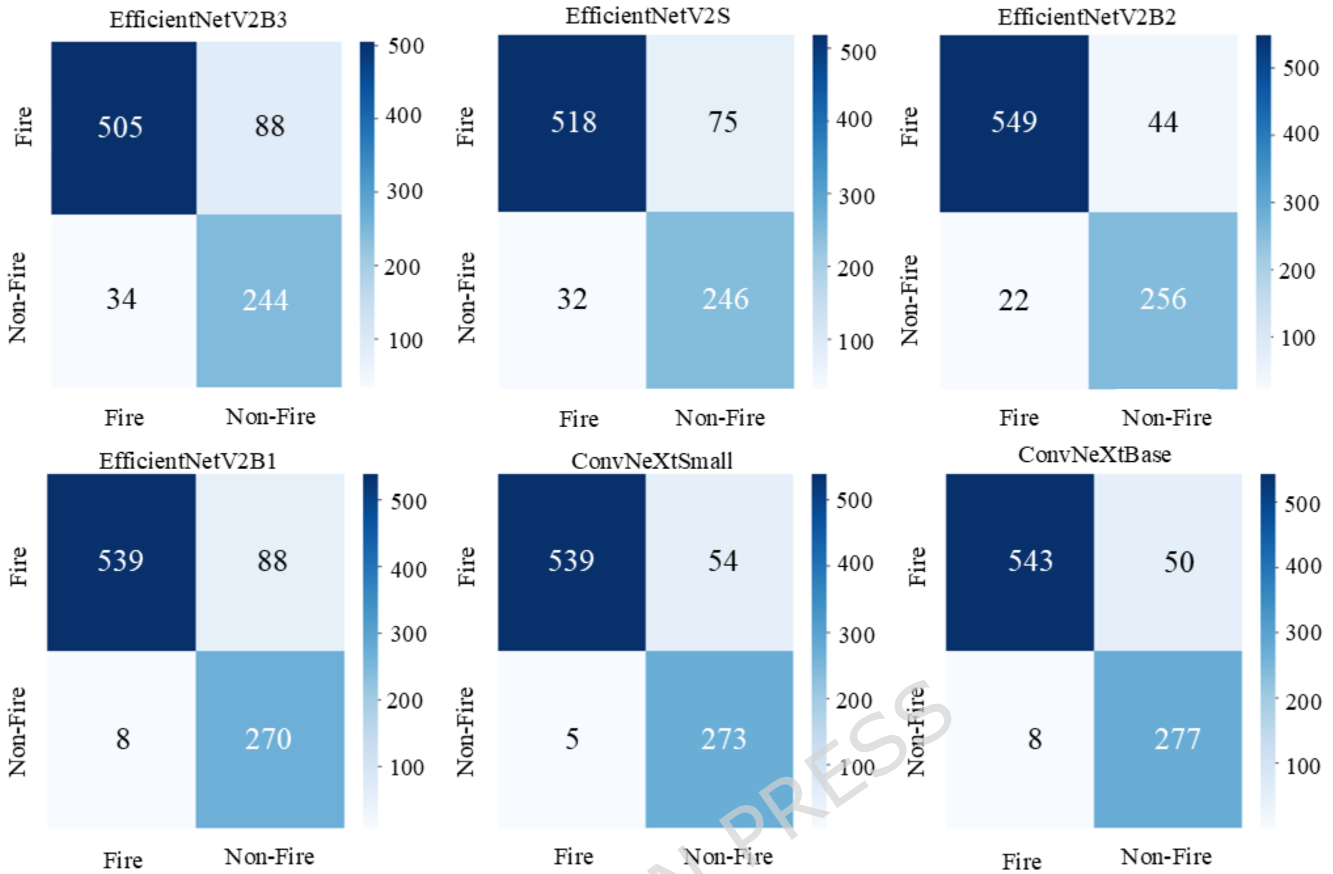

Figure 6. Performance evaluation of various CNN models using FireNet testing data.

Figure 7. Feature activation maps generated through the proposed model using Foggia, FireNet and DFV datasets.

5.2.2 Receiver Operating Characteristic (ROC) Curve, Accuracy and Loss Graphs Analysis

In binary prediction problems with balanced and imbalanced classes, the ROC can be used as a diagnostic tool because it is not biased against either the minority and the majority. It gives visual results that are convenient for researchers to understand the performance of the model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR). In this study, we produce ROC curves of three datasets of fire and non-fire, including Foggia's, Drone and FireNet.The ROC curves demonstrate the effectiveness and robustness of the models in distinguishing between fire and non-fire instances. A greater Area Under the Curve (AUC) value shows better discrimination ability. In the drone dataset, the fire classification model accomplished an AUC of 0.97, while the non-fire AUC was 0.96. In both the benchmarks' datasets, including FireNet and Foggia's datasets, the fire AUC increased to 100, with a non-fire AUC of 1.00, indicating excellent performance in classification because the proposed AFN network encodes both global and dependencies across channels and local spatial patterns associated with fire.

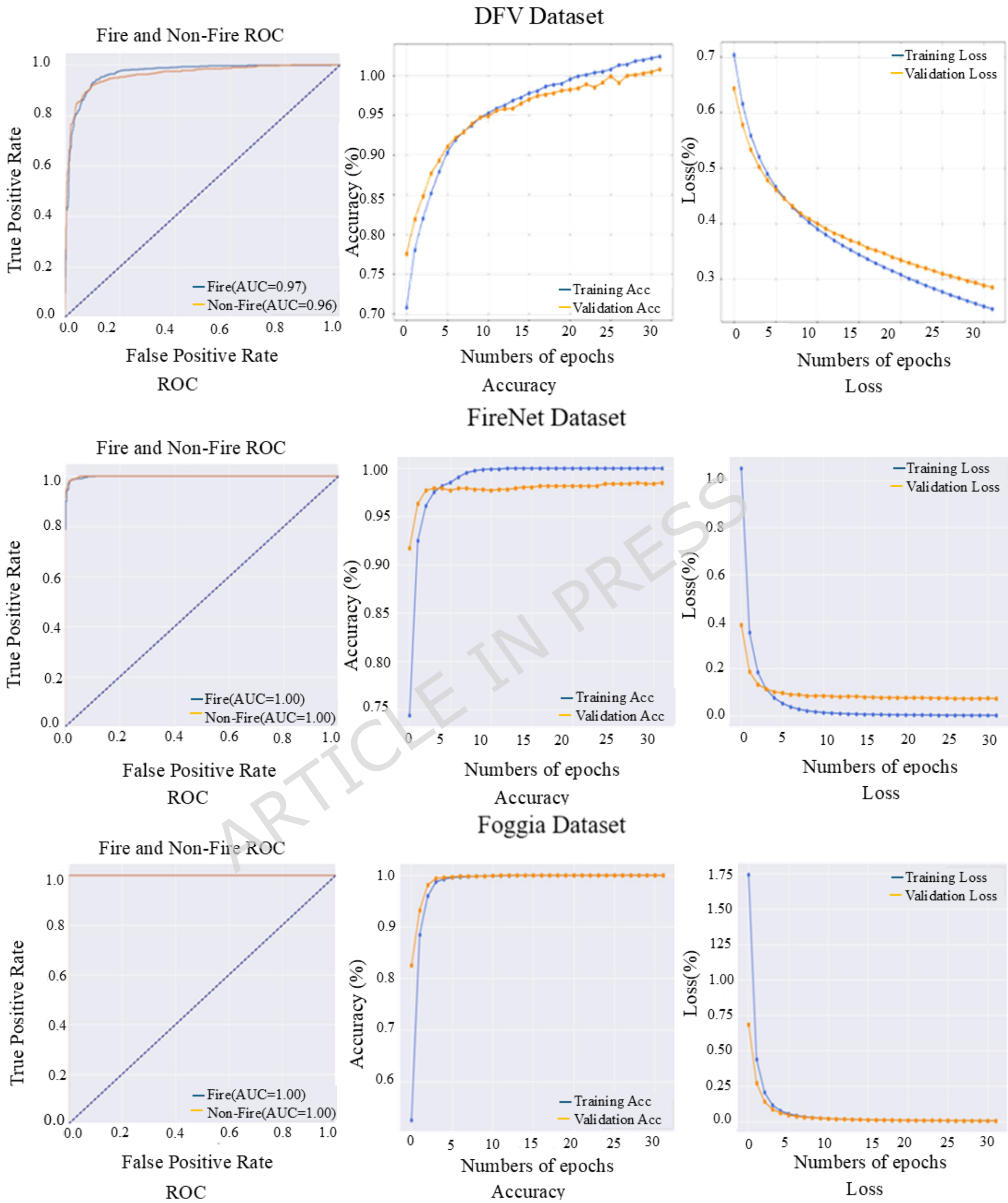

Figure 8. ROC Curves, Accuracy, and Loss graphs of AFN Network. The first row represents the ROC curve, Accuracy, and loss of the DFV dataset, while the second and third rows illustrate the ROC Curve, Accuracy, and Loss for FireNet and Foggia dataset, respectively.

Furthermore, they capture contextual information to differentiate between fire and non-fire. **Figure 8** illustrates that the overall performance of the proposed AFN network is convenient and demonstrates good performance for UAV-based fire detection and classification.

6 Computational Efficiency

In real-time fire monitoring systems, inference speed is a crucial factor because early fire recognition requires fast and reliable decision-making, especially in UAV-assisted and resource-constrained deployment environments. Therefore, the inference speed of the proposed AFN was compared with several existing fire detection methods in terms of GPU FPS and CPU FPS, as reported in Table 7. The results demonstrate that the proposed AFN achieves the highest inference speed among all compared methods, reaching 76.9 FPS on GPU and 23.6 FPS on CPU. Among the existing approaches, ADFireNet obtains the second-best performance with 72.5 GPU FPS and 22.0 CPU FPS, followed by DFAN with 70.55 GPU FPS and 12.90 CPU FPS. Although these methods demonstrate strong real-time capability, the proposed AFN further improves the inference speed by 4.4 FPS on GPU and 1.6 FPS on CPU compared with ADFireNet. Similarly, compared with DFAN, AFN improves GPU speed by 6.35 FPS and CPU speed by 10.7 FPS. This indicates that the proposed architecture not only maintains high classification performance but also provides efficient inference on both GPU- and CPU-based platforms.

Other methods, such as ResNetFire, GNetFire, EMNFire, SE-EFFNet, and ViT-B-32, show comparatively lower inference speed, particularly in CPU settings. For instance, ResNetFire achieves 57.3 FPS on GPU but only 2.4 FPS on CPU, while EMNFire and GNetFire obtain 6.3 FPS and 4.3 FPS on CPU, respectively. In contrast, AFN achieves 23.6 FPS on CPU, demonstrating its practical suitability for resource-limited environments where GPU acceleration may not always be available. Moreover, the higher GPU FPS of AFN confirms that the integration of SAAN with ConvNeXtBase does not introduce excessive computational overhead. Overall, the FPS comparison confirms that AFN is an inference-efficient fire classification framework. The proposed model achieves the highest throughput among the compared methods while maintaining strong classification performance, supporting its suitability for deployment-oriented fire monitoring applications.

Table 7. Comparison of inference speed with state-of-the-art fire detection methods.

Reference	GPU FPS	CPU FPS
ANetFire[43]	–	17.0
ResNetFire[59]	57.3	2.4
EMNFire[21]	34.0	6.3
GNetFire[44]	48.2	4.3
DFAN[36]	70.55	12.90
SE-EFFNet[50]	45.0	8.0
ViT-B-32[48]	67.0	20.0
ADFireNet[2]	72.5	22.0
AFN (Proposed)	**76.9**	**23.6**

7 Conclusion, Limitations, and Future Research Directions

Vision-based fire detection is increasingly important for surveillance and remote sensing. Despite recent advances in deep learning, reliable fire recognition remains challenging in complex scenes due to illumination changes, background clutter, variable fire appearance, and fire-like objects. To address these challenges, this study proposed the Adaptive Fire Network (AFN) for binary fire-scene classification. The proposed framework integrates a ConvNeXt-B backbone with the Scene-Aware Attention Network (SAAN), which consists of the Adaptive Channel Module (ACM) and the Adaptive Spatial Module (ASM). The ACM models dependencies among feature channels to emphasize informative fire-related responses, whereas the ASM highlights spatial regions that contribute to distinguishing fire from non-fire scenes. Depth-wise separable convolution is incorporated into the spatial-attention process to refine spatial features while limiting additional computational overhead. AFN was evaluated on the Foggia, FireNet, and newly introduced Drone Fire Vision (DFV) datasets. Under the reported experimental settings, the model achieved classification accuracies of 99.96% on Foggia, 98.97% on FireNet, and 90.78% on DFV. The ablation experiments showed that the complete configuration combining ACM and ASM achieved higher classification performance than the evaluated ConvNeXt-B baseline and configurations using partial or alternative attention mechanisms. These findings indicate that sequential channel-wise and spatial feature refinement can improve the representation of fire and non-fire scenes within the evaluated datasets. The reported inference rates of 76.9 frames per second on the evaluated GPU and 23.6 frames per second on the evaluated CPU also indicate efficient processing under the hardware and implementation conditions used in this study. Nevertheless, several limitations should be considered when interpreting these results. First, AFN performs image-level binary classification and therefore assigns a fire or non-fire label to the entire input image. It does not explicitly localize or segment flame and smoke regions, distinguish among multiple fire regions, estimate fire severity, or determine the burning degree or growth rate of a fire. Estimating such properties generally requires spatial localization, consecutive observations, and additional environmental or sensor information. In addition, the current framework focuses on visual classification and does not generate detailed semantic descriptions of the surrounding scene, such as the environmental context or potential sources of visual ambiguity. Moreover, there are several promising ways for future research. Firstly, we aim to explore the integration of advanced sensor technologies, such as thermal imaging, to enhance the model's capabilities in detecting fires in adverse weather conditions, such as dense fog or smoke. Secondly, we intend to investigate the integration of vision-language models to improve semantic scene understanding, distinguish fire from visually similar objects, and generate more informative outputs beyond binary classification. Furthermore, we plan to explore novel deep-learning architectures and optimization strategies to reduce computational complexity and memory requirements, thereby enhancing the scalability and effectiveness of the model for deployment on resource-constrained UAV platforms.

Funding

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the virtual convergence support program to nurture the best talents(IITP-2026-RS-2023-00254529) grant funded by the Korea government(MSIT) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (RS-2026-25473871).

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the virtual convergence support program to nurture the best talents(IITP-2026-RS-2023-00254529) grant funded by the Korea government(MSIT) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (RS-2026-25473871).

Conflicts of Interest

The authors declare no conflicts of interest.

7.1 Data availability

The study used the Firenet dataset is available at *https://github.com/arpit-jadon/FireNet-LightWeight-Network-for-Fire-Detection?tab=readme-ov-file* and the Foggia dataset, which is available at *https://mivia.unisa.it/datasets/video-analysis-datasets/fire-detection-dataset/*. An DFV dataset is available at *https://github.com/SufyanDanish/Drone-Fire-Vision/tree/mai*

Author contributions statement

Conceptualization, methodology, statistical analysis, data analysis: S.D; literature review, discussion, writing-original draft preparation, data downloading: S.D, H.M and S.U; writing-review and editing: L.M.D, H.M and H.K.S; visualization: S.D, L.M.D, and S.U; supervision: H.M All authors read and approved the final manuscript.

References

1. Danish, S. *et al.* Vision-based fire management system using autonomous unmanned aerial vehicles: a comprehensive survey. *Artif. Intell. Rev.* **59**, 16 (2025).

2. Yar, H., Ullah, W., Khan, Z. A. & Baik, S. W. An effective attention-based cnn model for fire detection in adverse weather conditions. *ISPRS J. Photogramm. Remote. Sens.* **206**, 335–346 (2023).

3. Brushlinsky, N., Ahrens, M., Sokolov, S. & Wagner, P. World fire statistics. *Cent. fire statistics.-2006.-Report* **10** (2016).

4. Margaret Stewart, L. A. F. D. Welcome to the los angeles fire department. https://lafd.org/news/palisades-fire-0 https://www.reuters.com/world/us/los-angeles-wildfires-rage-third-night-death-toll-rises-10-2025-01-10/ (2025). Accessed: 2025-01-10.

5. Wang, H., Jin, B., Zhang, K., Aktar, S. & Song, Z. Effectiveness in mitigating forest fire ignition sources: A statistical study based on fire occurrence data in china. *Fire* **5**, 215 (2022).
6. Xu, H. *et al.* Temperature-triggered sensitive resistance transition of graphene oxide wide-ribbons wrapped sponge for fire ultrafast detecting and early warning. *J. hazardous materials* **363**, 286–294 (2019).
7. Baek, J. *et al.* Intelligent multi-sensor detection system for monitoring indoor building fires. *IEEE Sensors J.* **21**, 27982–27992 (2021).
8. Gaur, A. *et al.* Fire sensing technologies: A review. *IEEE Sensors J.* **19**, 3191–3202 (2019).
9. Clark, H. Fire and rescue incident statistics: England, year ending december 2021 (2022).
10. Dilshad, N., Khan, S. U., Alghamdi, N. S., Taleb, T. & Song, J. Towards efficient fire detection in iot environment: A modified attention network and large-scale dataset. *IEEE Internet Things J.* (2023).
11. Chen, T.-H., Wu, P.-H. & Chiou, Y.-C. An early fire-detection method based on image processing. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, vol. 3, 1707–1710 (IEEE, 2004).
12. Nakıp, M., Keleşoğlu, N. & Güzeliş, C. Fire detection and risk assessment via support vector regression with flattening-samples based augmented regularization. *Appl. Soft Comput.* **164**, 112023 (2024).
13. Zhou, X. *et al.* Characterizing fire and fire atmospheric states from space using collocated hyperspectral infrared sounding and narrow-band imagery. *Remote. Sens. Environ.* **312**, 114318 (2024).
14. Danish, S. *et al.* Hqdl: A hybrid quantum-deep learning framework for real-time fire detection using uav remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* (2026).
15. Woodgate, W., Phinn, S., Devereux, T. & Aryal, R. R. Bushfire recovery at a long-term tall eucalypt flux site through the lens of a satellite combining multi-scale data for structural-functional insight. *Remote. Sens. Environ.* **317**, 114530 (2025).
16. Wang, S. *et al.* Hierarchical scale awareness for object detection in unmanned aerial vehicle scenes. *Appl. Soft Comput.* **168**, 112487 (2025).
17. Debnath, S., Ahmed, S., Das, S., Nahid, A.-A. & Bairagi, A. K. Iot based low-cost gas leakage, fire, and temperature detection system with call facilities. In *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, 11–16 (IEEE, 2020).
18. Cheon, J. *et al.* A single-chip cmos smoke and temperature sensor for an intelligent fire detector. *IEEE Sensors J.* **9**, 914–921 (2009).
19. Celik, T. & Demirel, H. Fire detection in video sequences using a generic color model. *Fire safety journal* **44**, 147–158 (2009).
20. Yu, C., Mei, Z. & Zhang, X. A real-time video fire flame and smoke detection algorithm. *Procedia Eng.* **62**, 891–898 (2013).
21. Muhammad, K., Khan, S., Elhoseny, M., Ahmed, S. H. & Baik, S. W. Efficient fire detection for uncertain surveillance environment. *IEEE Transactions on Ind. Informatics* **15**, 3113–3122 (2019).
22. Khan, S. U. *et al.* Firemamba: An efficient visual representation learning framework for fire detection. *Alex. Eng. J.* **132**, 133–144 (2025).
23. Dilshad, N. *et al.* Yoloflame: Benchmarking vehicle fire and smoke detection in urban environments. *Alex. Eng. J.* **141**, 109–118 (2026).
24. Chen, L.-H., Kolhe, S. S., Hu, J., Tseng, K.-H. & Chung, M.-Y. An integrated edge–cloud iot framework for resilient disaster prevention in fire detection and forest carbon assessment. *Sci. Reports* (2026).
25. Abozeid, A. & Alanazi, R. An intelligent approach for early smoke/fire detection using vision sensors in smart cities. *Sci. Reports* (2026).
26. Wu, W., Zhou, X., Qin, J., Fu, Z. & Xing, K. Research on target detection algorithm for forest fire images based on multi-scale feature extraction. *Sci. Reports* (2026).
27. Liu, X., Zhang, Q., Zhang, Y. & Zhu, J. Enhancing deep learning based multi-parameter fire detection by training on simulated and combustion experimental dataset. *Sci. Reports* (2026).
28. Lilhore, U. K. *et al.* Real time fire and smoke detection using vision transformers and spatiotemporal learning. *Sci. Reports* (2026).

29. Ul Haq, I., Husnain, G., Iqbal, A. & Alzahrani, A. S. Attention-enhanced mobilenetv2 models for robust forest fire detection and classification. *Sci. Reports* (2026).

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

31. Saeed, F., Paul, A., Karthigaikumar, P. & Nayyar, A. Convolutional neural network based early fire detection. *Multimed. Tools Appl.* **79**, 9083–9099 (2020).

32. Saeed, F., Paul, A., Karthigaikumar, P. & Nayyar, A. Convolutional neural network based early fire detection. *Multimed. Tools Appl.* **79**, 9083–9099 (2020).

33. Khudayberdiev, O., Butt, M. H. F. *et al.* Fire detection in surveillance videos using a combination with pca and cnn. *Acad. J. Comput. Inf. Sci* **3**, 27–33 (2020).

34. Danish, S., Khan, S. U., Dang, L. M., Song, H.-K. & Moon, H. Toward intelligent earth observation: Hierarchical feature fusion and a drone dataset for uav-based fire detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* (2026).

35. Guan, Z. *et al.* Forest fire segmentation from aerial imagery data using an improved instance segmentation model. *Remote. Sens.* **14**, 3159 (2022).

36. Yar, H. *et al.* Optimized dual fire attention network and medium-scale fire classification benchmark. *IEEE Transactions on Image Process.* **31**, 6331–6343 (2022).

37. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).

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

39. Jadon, A., Omama, M., Varshney, A., Ansari, M. S. & Sharma, R. Firenet: a specialized lightweight fire & smoke detection model for real-time iot applications. *arXiv preprint arXiv:1905.11922* (2019).

40. Du, D. *et al.* Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0 (2019).

41. Rafiee, A., Dianat, R., Jamshidi, M., Tavakoli, R. & Abbaspour, S. Fire and smoke detection using wavelet analysis and disorder characteristics. In *2011 3rd International conference on computer research and development*, vol. 3, 262–265 (IEEE, 2011).

42. Di Lascio, R., Greco, A., Saggese, A. & Vento, M. Improving fire detection reliability by a combination of videoanalytics. In *Image Analysis and Recognition: 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part I 11*, 477–484 (Springer, 2014).

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

44. Muhammad, K., Ahmad, J., Mehmood, I., Rho, S. & Baik, S. W. Convolutional neural networks based fire detection in surveillance videos. *Ieee Access* **6**, 18174–18183 (2018).

45. Muhammad, K. *et al.* Efficient deep cnn-based fire detection and localization in video surveillance applications. *IEEE Transactions on Syst. Man, Cybern. Syst.* **49**, 1419–1434 (2018).

46. Hashemzadeh, M. & Zademehdi, A. Fire detection for video surveillance applications using ica k-medoids-based color model and efficient spatio-temporal visual features. *Expert. Syst. with Appl.* **130**, 60–78 (2019).

47. Jadon, A., Varshney, A. & Ansari, M. S. Low-complexity high-performance deep learning model for real-time low-cost embedded fire detection systems. *Procedia Comput. Sci.* **171**, 418–426 (2020).

48. Shahid, M. & Hua, K.-L. Fire detection using transformer network. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 627–630 (2021).

49. Li, Y., Zhang, W., Liu, Y. & Jin, Y. A visualized fire detection method based on convolutional neural network beyond anchor. *Appl. Intell.* **52**, 13280–13295 (2022).

50. Khan, Z. A. *et al.* Randomly initialized cnn with densely connected stacked autoencoder for efficient fire detection. *Eng. Appl. Artif. Intell.* **116**, 105403 (2022).

51. Avula, S. B., Badri, S. J. & Reddy, G. A novel forest fire detection system using fuzzy entropy optimized thresholding and stn-based cnn. In *2020 international conference on COMmunication Systems & NETworkS (COMSNETS)*, 750–755 (IEEE, 2020).

52. Khan, Z. A. *et al.* Optimized cross-module attention network and medium-scale dataset for effective fire detection. *Pattern Recognit.* **161**, 111273 (2025).

53. Gongsuo, C. *et al.* Fasternetfire: A cost-effective fast neural network for forest fire detection with partial convolution. *Forests* **17**, 672 (2026).

54. Xie, K. *et al.* A convolution-augmented swin transformer for robust fire detection with enhanced anti-interference capability. *Appl. Soft Comput.* 115429 (2026).

55. Nguyen, D.-L., Putro, M. D. & Jo, K.-H. Lightweight convolutional neural network for fire classification in surveillance system. *IEEE Access* (2023).

56. Hosseini, A., Hashemzadeh, M. & Farajzadeh, N. Ufs-net: A unified flame and smoke detection method for early detection of fire in video surveillance applications using cnns. *J. Comput. Sci.* **61**, 101638 (2022).

57. Nguyen, D.-L., Putro, M. D., Vo, X.-T., Tran, T.-D. & Jo, K.-H. Fire warning based on convolutional neural network and inception mechanism. In *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 1–5 (IEEE, 2022).

58. Habiboğlu, Y. H., Günay, O. & Çetin, A. E. Covariance matrix-based fire and flame detection method in video. *Mach. Vis. Appl.* **23**, 1103–1113 (2012).

59. Sharma, J., Granmo, O.-C., Goodwin, M. & Fidje, J. T. Deep convolutional neural networks for fire detection in images. In *International conference on engineering applications of neural networks*, 183–193 (Springer, 2017).