Toward Intelligent Earth Observation: Hierarchical Feature Fusion and a Drone Dataset for UAV-Based Fire Detection

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

***Abstract*—In UAV-based monitoring and remote sensing, early fire detection remains challenging due to the complex appearance, scale variation, and spatial ambiguity of fire scenes. Existing deep networks often rely heavily on feature maps from preceding layers, which can weaken fine-grained spatial representation and reduce detection accuracy, particularly for small, distant, or irregular fire regions. To address these limitations, we propose a Hierarchical Feature Fusion (HFF) framework that captures fire patterns across multiple scales. The proposed framework integrates Discriminative Fire Features (DFF) with a Fire Guided-Attention (FGA) module to refine hierarchical representations and improve sensitivity to varying fire intensities, distances, and scene conditions. In addition, to address the limited diversity of existing datasets, we introduce a new Drone Fire (DF) dataset containing a broad range of fire scenarios, geographic settings, and environmental conditions, while reducing class imbalance. Extensive experiments on four datasets, namely DF, FD, FLAME, and DSFD, show that HFF achieves accuracies of 89.16%, 97.52%, 97.50%, and 96.00%, respectively. The results demonstrate that the proposed framework provides reliable and robust performance across diverse UAV-based fire monitoring scenarios, making it well suited for remote sensing applications in fire surveillance and disaster management.**

***Index Terms*—Fire detection, UAV remote sensing, disaster monitoring, deep learning, hierarchical feature fusion, UAV dataset.**

I. Introduction

Fire is one of the most destructive environmental hazards, demanding rapid and reliable detection to reduce losses to ecosystems, infrastructure, and human life. Among natural and human-induced hazards, fire is particularly difficult to manage due to its rapid spread, evolving visual characteristics, and the limited time available for intervention. The destructive nature of wildfires and bushfires is particularly harmful as geographic spreads are usually big and the loss is devastating in terms of ecological and economic damage. The recent massive events are another indication that there is a necessity of having effective early warning systems. For example, a major fire in Los Angeles in 2025, where over 5,316 buildings were destroyed and the area of about 13,750 hectares was burned, and in California, in 2024, there was a greater number of 2,429 fire incidents, and over 95,562 acres were burned and more than 50 buildings were destroyed. In the European Union, about every 87 seconds in 2024 a home fire incident was reported [1]. These incidents highlight the real-world significance of accurate and timely fire detection.

The methods of the traditional fire monitoring system are primarily based on the principle of scalar sensors, including temperature, flame, smoke, and particle sensors [2]. Even though these systems are inexpensive and simple to implement, they are usually restricted to closed systems, offer localized measurements, and in practice, may need further validation. Conversely, vision-based systems can cover a broader range, respond faster, and provide more detailed scene information, making them better suited for disaster monitoring and environmental analysis. As UAVs and other imaging systems gain more and more popularity, vision-based fire detection has become a scalable system that can be employed to survey large-scale areas [3]. The current vision-based methods can be broadly divided into traditional machine learning (TML), hybrid methods that combine TML and deep learning (DL), and end-to-end DL-based approaches.

Initial TML-based fire detection algorithms used handcrafted features, including color, texture, motion, and shape, to differentiate between fire and background areas [4]. Color-based methods usually used RGB or YCbCr coding to determine fire-like pixels, whereas other schemes used color and texture features, fuzzy logic, covariance features, or support vector machines to enhance the performance of the classification process [5]–[9] These techniques are usually sensitive to variations in illumination, smoke, shadow, messy backgrounds and objects that resemble fire, producing high false-positive rates and erratic operation. More significantly, handcrafted elements are limited in their ability to capture the broad diversity of fire in terms of scale, perspective, and contextual circumstances [10].

To overcome these restrictions, recent research has been progressively using DL-based solutions. CNNs learn to represent discriminative information directly with image data and have demonstrated explicit benefits over handcrafted fire detection techniques [11], [12]. Existing work has explored lightweight architectures, model compression, transfer learning, transformer-enhanced designs, and hybrid CNN-based

S. Danish, and H. Moon are with the Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea (e-mail: sufyandanish@sju.ac.kr; hmoon@sejong.ac.kr).S. U. Khan with the Advanced Research and Innovation Center, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates (e-mail: samee.khan@ku.ac.ae).L. Minh Dang is with the Institute of Research and Development and the Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam and the Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea H.K. Song is with the Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea (e-mail:songhk@sejong.ac.kr). Manuscript received April 16, 2026 (Corresponding authors: Hyeonjoon Moon). 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 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 "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government. (2026-RISE-01-019-04).

frameworks combined with conventional classifiers such as support vector machines [13]–[20]. Beyond direct fire recognition, DL methods have also been applied to related wildfire tasks, including air-quality impact analysis, thermal hotspot monitoring, fire severity assessment, burned-area analysis, and drone-assisted fire management [3], [21]–[28]. Fig. 1 given Supplementary file present the a graphical representation of fire-related work categorized into hybrid, machine learning and deep learning methods.

Despite this progress, important limitations remain. Many DL models are trained and evaluated on datasets collected from fixed surveillance cameras, which provide limited viewpoint diversity and restricted spatial coverage, thereby weakening generalization to open and dynamic outdoor environments. Moreover, the area of fire can be highly diverse in size, shape, texture, and intensity, and small or remote fires are especially hard to notice. Fixed feature extraction pipelines might also be fail in extracting fine detail of local features as well as more general contextual information of complex fire scenes. Despite the advancement on the matter of feature selectivity, attention-based approaches have significant shortcomings. Bottleneck attention can over-compress information, self-attention can be costly and inefficient to maintain local structure, and channel-spatial attention like CBAM can fail to represent hierarchical dependencies across scale ranges in a sufficient manner, among others [11], [29]–[31]. Attention-guided P-DenseNet-A-TL and channel- or semantic-attention-based frameworks have demonstrated positive improvements in attention-guided fire detection, although they are so far entirely reliant on sequence backbone feature maps, which may compromise local spatial information especially with early-stage or low-sized fires and in scenes containing fire-like distractions or high-contrast illumination variations.

Another practical challenge is a shortage of sufficiently diverse datasets to monitor fire aerospace. The available benchmarks usually lack sufficient drone-based images to capture the variability of scale, perspective shifts, and complexity of the environment typical of outdoor real-world monitoring. Aerial imagery has the potential to offer wider spatial coverage and more versatile observation geometry compared to fixed CCTV systems, and is specifically useful in fire surveillance of large, heterogeneous areas. However, the limited availability of such data continues to constrain the development and fair evaluation of robust fire detection models.

Motivated by these observations, this work proposes a Hierarchical Feature Fusion (HFF) framework for fire scene classification in diverse visual sensing conditions. The proposed framework is designed to improve robustness to scale variation, background complexity, and ambiguous visual patterns by integrating hierarchical representations across multiple resolutions. In contrast to approaches that rely on a single feature level, HFF aggregates complementary information from different stages of the network, enabling more reliable modeling of both small-scale and large-scale fire characteristics. To further enhance discriminative capability, a Fire Guided-Attention (FGA) module is introduced to emphasize spatially informative regions and to refine hierarchical features using both local and global contextual dependencies. This design improves resilience to fire-like distractors and to illumination variability, as illustrated in Fig. 2 of the Supplementary.

To support this study, we also construct a new drone-based fire dataset, denoted as the DF dataset, consisting of 1,130 fire images and 1,130 non-fire images captured under challenging conditions. The dataset is intended to address the shortage of aerial fire samples and to provide improved geographic and visual diversity for model evaluation. Extensive experiments are conducted on the DF, FLAME, DSFD, and FD datasets using multiple backbone architectures and several variants of the proposed FGA design. Among the examined backbones, EfficientNetB2 provides the most favorable performance for the proposed framework. In the final architecture, EfficientNetB2 is used as the feature extractor, while the proposed FGA module refines hierarchical features prior to fusion and classification.

Although UAV-based imagery is the primary focus of this work, the FD dataset was additionally included to evaluate the generalization capability of the proposed framework under mixed visual sensing conditions. Specifically, DF, FLAME, and DSFD mainly consist of drone/aerial imagery, whereas FD contains both drone-based and CCTV/ground-based fire images. Therefore, FD is used as a cross-perspective benchmark rather than the primary target domain.

We also re-implemented recent state-of-the-art fire detection models and evaluated them on the DF dataset, as illustrated in Fig. 2 of the Supplementary. The comparative results show that several existing methods remain susceptible to misclassification in challenging outdoor scenes, particularly when fire-like objects or difficult visual conditions are present. By contrast, the proposed HFF model provides more reliable classification on these challenging samples, demonstrating the importance of hierarchical feature integration and fire-aware attention for robust fire scene analysis.

The main contributions of this work are summarized as follows:

- We propose a hierarchical feature fusion framework, termed HFF, primarily designed for UAV-based fire scene classification while also demonstrating robust generalization across heterogeneous visual sensing conditions. By combining multi-level contextual and spatial representations, the framework improves discrimination of complex fire patterns under diverse scene conditions.
- We introduce a hierarchical Fire Guided-Attention module that enhances feature refinement across multiple scales. The proposed module preserves local spatial information while incorporating broader contextual dependencies, thereby improving robustness to scale variation, background clutter, and visually similar non-fire regions.
- We develop a new drone-based fire dataset, referred to as DF, to address the limited availability of challenging aerial fire imagery. Extensive experiments on DF and FD demonstrate that the proposed HFF framework outperforms traditional handcrafted-feature methods, conventional CNN-based baselines, and existing attention-based approaches.

The remainder of this paper is organized as follows. Section II describes the proposed HFF architecture in detail. Section

III presents the datasets, evaluation metrics, and experimental settings. Section IV reports comparative results against state-of-the-art methods, followed by further analysis in Section V. VI discuss computational complexity. Finally, Section VII concludes the paper and outlines future research directions.

II. Proposed Framework

This section describes a comprehensive overview of the proposed HFF framework, which comprises three major components: DFF, FGA and HFF. The overall workflow comprises three sequential phases: data preprocessing, training and testing. In the data preprocessing phase, video data are collected from various sources and converted into frames. In both the training and testing stages, a number of CNN-based models were implemented to perform a thorough analysis. Among them, EfficientNetB2 proved to be the most effective and was chosen as the DFF, which should capture detailed, fire-specific features that are important in the fine-grained fire detection. To overcome the scale variation and contextual sensitivity, Hierarchical fire-guided-attention mechanism was integrated, with multi-kernel convolutions to add sensitivity to a range of fire sizes. The output feature map of each of the two modules was then incorporated through a hierarchical strategy of feature fusion, with this design enabling the model to pay attention to both localized and large-scale fires, as well as to a greater number of possible fire scenarios, and enables the model to have a better classification and localization accuracy. The comprehensive framework of the proposed model is presented in Fig. 1. The details of each phase, including the additional technical details, are also provided in the sub-sub sections below.

A. Overview

Fire poses a serious threat to ecological and environmental systems, which is increased by the growing impacts of climate change. Fire disasters can cause severe environmental pollution, destruction of natural resources, loss of human lives in large numbers and major economic damage. Therefore, early and accurate fire detection is essential, and several ML and DL-based models have been developed for this purpose. However, many existing methods remain sensitive to scale variation and depend heavily on preceding feature layers, which may limit their ability to capture discriminative spatial details. To address these limitations, this work proposes a novel Hierarchical Feature Fusion (HFF) framework for fire detection and classification from drone images. The proposed framework consists of three main phases, including data preprocessing, training, and testing. In the preprocessing phase, frames are extracted from multimedia data collected from different sources. During training, the fire dataset is first passed through the Discriminative Fire Feature ($\mathfrak{DFF}$) model to extract representative fire-specific features. These features are then processed by the Fire Guided-Attention (FGA) module through the functions F_x, G_x, and H_x, enabling the model to capture both fine-grained and broad fire patterns. In parallel, the extracted features are also processed through $\mathcal{J}_x$ and $\mathcal{K}_x$. The hierarchical features obtained from the FGA module are then fused with the backbone features to generate a more informative final descriptor. The final classification is performed using batch normalization, average pooling, fully connected layers with two neurons, and a SoftMax function to distinguish between fire and non-fire classes. After training, the model is saved and later loaded during the testing phase to evaluate its performance using accuracy, precision, and F1-score. The overall workflow of the proposed framework is summarized in Algorithm 1 in the Supplementary Material, and further details are provided in the following sections.

B. Pre-Processing

To overcome the issues of classification and detection in the fire detection domain, it is essential to have a systematically collected and organized diverse dataset of fire and non-fire cases. The dataset has been collected based on 66 multimedia data sources across various online sources, such as Facebook, YouTube, and Instagram and specific keywords, such as 'fire', 'building fire', 'forest fire', 'vehicle fire' and 'mountains fire'. We have also used the words, including 'drone', 'UAV' and 'Aerial' in our search. This multimedia data was then converted into frames and a total of 6000 drone fire frames were obtained in all the multimedia data.

Ensuring the quality of our dataset is critical for effective model training and optimal results. To achieve this, several key preprocessing actions were carried out in which each image was visually examined to retain only unique and relevant data and discard duplicated, blur, non-informative, low-quality and redundant data to narrow the range of drone fire images to 1130 but discard unnecessary and redundant frames to avoid model overfitting. This data set is consists of diverse types of data such as geographic and environmental diversity. Each image was carefully examined, and only the relevant region of interest was retained, deliberately excluding irrelevant objects. Following this, all images were passed through image processing and resized to a standardized resolution of 224×224, facilitating the algorithm in efficiently learning features for the classification problem.

C. Fire Specific Feature Extraction

In fire detection tasks, specifically drone-based imagery with complex environmental contexts, it is critical to select a robust feature extraction backbone. To identify an optimal backbone for robust feature selection and pattern recognition in fire detection and classification tasks, especially in complicated scenarios, nine feature extractors, including InceptionV3, DenseNet121, MobileNet, MobileNetV2, NASNet-Mobile, EfficientNetB0, EfficientNetB1, EfficientNetB2 and EfficientNetB3 were comparatively evaluated based on their ability to extract semantically rich and spatially discriminative fire features. As shown in the ablation study (**Table I**), EfficientNetB2 consistently achieved the highest performance across classification accuracy and F1-Score metrics. Based on empirical validation, the EfficientNetB2 network was selected as a DFF for Fire-Specific Feature Extraction within the proposed HFF framework.

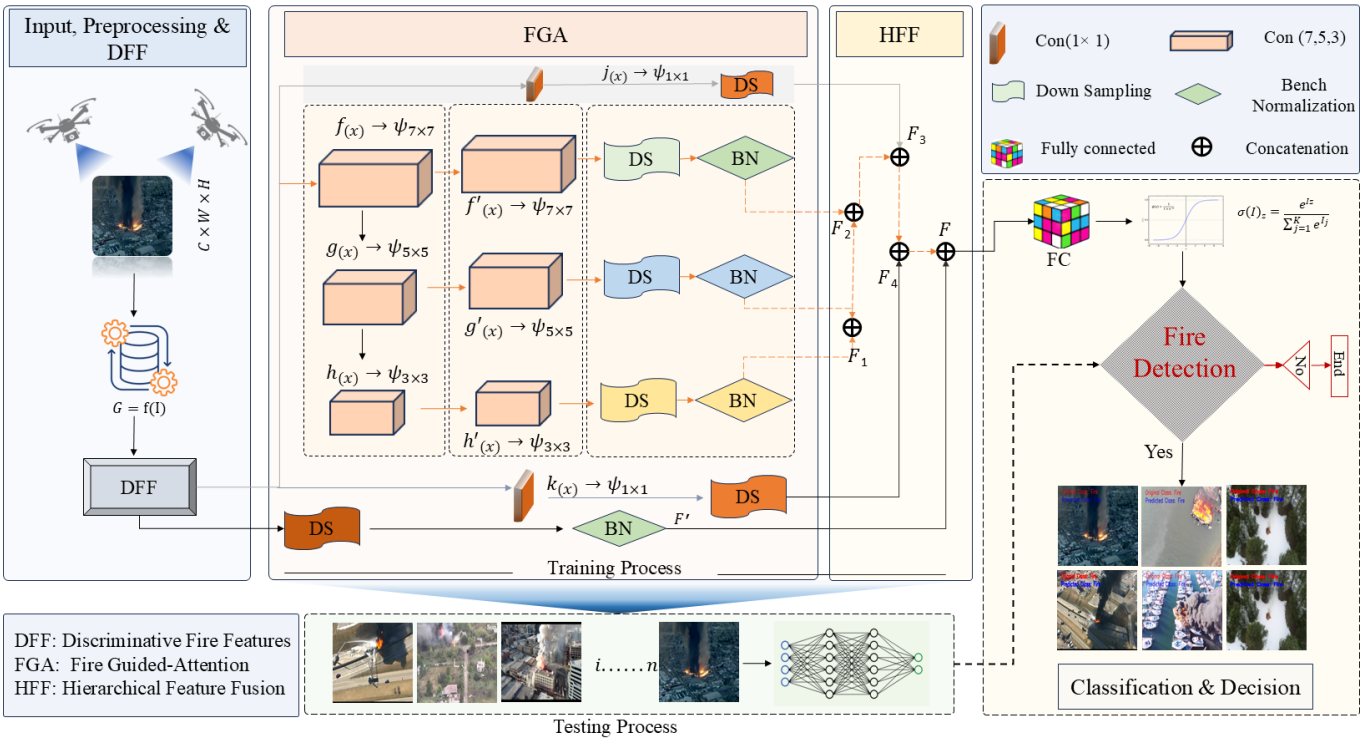

Fig. 1. Overall architecture of the proposed HFF framework. UAV fire images are first preprocessed and passed to the DFF module for discriminative feature extraction. The extracted feature maps are then refined using the proposed Fire Guided-Attention (FGA) module, which employs hierarchical multiscale attention to emphasize fire-relevant spatial regions. Subsequently, the refined features are integrated through the Hierarchical Feature Fusion (HFF) module to combine local and global contextual information to obtain the best output.

Beyond empirical superiority, EfficientNetB2 presents several architectural advantages that make it uniquely suitable for UAV-based fire detection tasks. In contrast to traditional CNNs where model parameters are scaled arbitrarily, EfficientNetB2 employs a compound scaling policy that scales depth, width, and resolution with constant coefficients, and thus reaches an optimal trade-off between accuracy and computational efficiency. This is especially beneficial to UAV platforms that lack processing power. Besides, EfficientNetB2 uses depthwise separable convolutions of depths 3×3 and 5×5, as well as squeeze-and-excitation (SE) blocks, which dynamically recalibrate channel-wise feature responses, improving the model to focus on spatially relevant patterns of fire on complex backgrounds. Gradient flow and model convergence is further enhanced by the swish activation function which allows more accurate representation learning of complex fire contours than traditional ReLU-based networks. Moreover, higher input resolutions support (224×224) in EfficientNetB2, which allows fine-grained spatial features to be preserved in the initial convolutional layers, which is essential to support fine-grained fire patterns in small scales and partially occluded fire patterns in the initial convolutional layers as well as scale variation in complex fire detection tasks. EfficientNetB2 is more advantageous in terms of performance to complexity compared to more advanced models like EfficientNetB3 and DenseNet121, which risk overfitting and high memory utilization. Although MobileNet and NASNetMobile are computationally efficient and lightweight, they tend to have a small ability to capture global context, which is necessary in the modeling of dispersed or ambiguous fire signals in aerial images. InceptionV3's use of asymmetric convolutions improves computational speed but compromises sensitivity to spatial structure, and DenseNet121, though effective in gradient propagation, tends to introduce redundant features and higher memory overhead. Furthermore, EfficientNetB2 offers a balanced architecture based on the compound scaling principle, where depth $\mathfrak{d}$, width $\mathcal{W}$, and resolution $\mathfrak{R}$ of the networks are scaled uniformly using a set of fixed coefficients ϕ. This relationship is defined mathematically as

$$\mathcal{E} = \quad \mathfrak{d} \propto \alpha^{\phi}, \quad \mathcal{W} \propto \beta^{\phi}, \quad \mathfrak{R} \propto \gamma^{\phi} \tag{1}$$

where $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ refer to the distribution of memory and computational resources towards the network width, depth and resolution respectively. This design enables EfficientNetB2 to maintain high accuracy while preserving computational efficiency a requirement for UAV-based fire detection applications.

The fire-specific feature extraction begins by processing the input image $I \in \mathbb{R}^{H \times W \times 3}$, which is resized to match the input dimensions required by the network (i.e., $224 \times 224 \times 3$). The image is subsequently passed through a sequence of convolutional and activation layers within EfficientNetB2, where low-level features such as edges and textures are extracted in the early stages. As the network depth increases, progressively more complex and abstract representations such as shapes, structural patterns, and semantic information are learned. The fire-specific feature extraction process can be expressed math-

ematically as follow:

$$FM = \mathcal{E}(I) \tag{2}$$

where $\mathcal{E}$ represents the EfficientNetB2 model, and FM represents the extracted fire-relevant feature maps from the input fire image I. These extracted features serve as the foundational input for subsequent modules in our framework. The strength of EfficientNetB2 in accurately capturing both high-level semantic context and low-level visual patterns is a crucial enabler of the strong fire detection performance of the system, as validated in the ablation study section III-A1 **Table I**, EfficientNetB2 consistently outperforms other evaluated backbones in terms of classification accuracy, F1-Score and other validation matrices, supporting its selection as the most suitable architecture for fire-specific feature extraction in UAV-based monitoring systems. However, the model does not rely exclusively on this feature map. Instead, the FGA and HFF modules are specifically designed to optimize, reweight, and add to these original properties through spatial hierarchies to reduce reliance on a single layer and make fire detection more generalizable in a variety of environmental conditions.

D. Hierarchical Fire Guided-Attention Mechanism

Conventional CNNs extract features within fixed receptive fields, which can restrict their ability to preserve fine spatial details, especially in shallow layers. In UAV-based fire detection, this limitation is critical because true fire regions often appear at different scales and may be confused with fire-like objects such as sunlight, reflections, smoke, or red-colored backgrounds. Although hierarchical structures improve multi-scale feature extraction, they may still lack explicit global contextual guidance. Likewise, channel-wise attention alone is insufficient because it does not fully preserve pixel-level spatial information. To address these challenges, we introduce an FGA module that improves the Discriminative Fire Feature by incorporating spatial hierarchies, multiscale contextual details, and attention-guided refinement, directly addressing the challenges of scale variation and overreliance on deep semantic maps highlighted in existing fire detection studies. The FGA module receives as input the fire-specific feature map extracted from DFF as denoted by $FM \in \mathbb{R}^{H\times W\times C}$. Initially, these features are normalized using batch normalization $\mathfrak{B}$ and passed through a dropout $\mathfrak{D}$ layer to improve generalization:

$$\hat{\mathbf{F}}_0 = \mathfrak{D}(\mathfrak{B}(\mathbf{FM})) \tag{3}$$

To capture fire patterns at multiple spatial resolutions, three convolutional layers with kernel sizes of 7×7, 5×5, and 3×3 are sequentially applied to $\hat{\mathbf{F}}_0$:

$$\mathbf{F} = \sigma\left(W_3 * \sigma\left(W_2 * \sigma\left(W_1 * \hat{\mathbf{F}}_0\right)\right)\right) \tag{4}$$

where σ denotes the ReLU activation function, W_i represents learnable convolutional kernels, and $*$ denotes convolution. The 7×7 and 5×5 kernels capture broad contextual fire structures, while the 3×3 kernel preserves local and fine-grained flame details. Since these operations are applied to high-level feature maps with reduced spatial dimensions, the use of larger kernels introduces only limited computational overhead. To enhance feature diversity, parallel attention branches are introduced at each hierarchical stage. For each intermediate feature map $\mathbf{F}_i$, an additional convolution with the corresponding kernel size is followed by global average pooling $\mathcal{G}_{ap}$ and batch normalization:

$$\mathbf{G}_i = \mathfrak{B}\left(\mathcal{G}_{ap}\left(\mathrm{Conv}_{k_i\times k_i}(\mathbf{F}_i)\right)\right) \quad i \in \{1,2,3\} \tag{5}$$

where $k_i \in \{7,5,3\}$. The resulting descriptors $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$ are concatenated to form a unified multi-scale attention representation:

$$\mathbf{F}_{ms} = \mathbf{G}_1 \oplus \mathbf{G}_2 \oplus \mathbf{G}_3 \tag{6}$$

where $\oplus$ denotes channel-wise concatenation. In addition, two 1×1 convolutional projections, $\mathbf{F}^{(1)}_{1\times 1}$ and $\mathbf{F}^{(2)}_{1\times 1}$, are extracted from the fire-specific feature map to retain compact channel-level information. A global average pooling descriptor $\mathbf{x}$ is also included to incorporate global contextual information. These branches ensure that global context is captured and added to the localized features, and further enhances the performance of the proposed FGA module. The final fire-refined feature representation is defined as:

$$\mathbf{F}_{\mathrm{FGA}} = \mathrm{Concat}\left(\mathbf{F}_{ms}, \mathbf{F}^{(1)}_{1\times 1}, \mathbf{F}^{(2)}_{1\times 1}, \mathbf{x}\right) \tag{7}$$

The FGA mechanism can also be interpreted as a weighted aggregation of multi-scale feature maps:

$$\mathbf{F}_{\mathrm{FGA}} = \sum_{i=1}^{n} \alpha_i \mathbf{FM}_{s_i} \tag{8}$$

where $\mathbf{FM}_{s_i}$ denotes the feature map from scale $s_i \in S$, and $S = \{s_1, s_2, \ldots, s_n\}$ represents the set of extracted scales. The attention coefficients α_i are normalized using a softmax function so that the contribution of each scale is adaptively weighted. This allows the module to emphasize the most informative fire-related regions while suppressing irrelevant background responses.

The other operations within the FGA module can be represented formally as follows: Equations **9**, **10**, **11**, **12**, **13** and **14** represent the convolutional layer operation, global average pooling, batch normalization, fully connected layer, loss function, activation function and loss function.

Convolutional Layer: The fully connected layers take the flattened feature representations as input and perform nonlinear transformations to discover intricate patterns within data. Convolutional layer with input tensor $\mathfrak{X}$, filter weights $\mathcal{W}$, bias $\mathfrak{B}$, and activation function σ, the convolution operation can be written as:

$$\mathrm{Conv}\quad 2\mathrm{D}(\mathfrak{X}, \mathcal{W}, \mathfrak{B}) = \sigma\left(\sum_{i,j} \mathfrak{X}_{i,j} * \mathcal{W}_{i,j} + \mathfrak{B}\right) \tag{9}$$

Here, * denotes the convolution operation.

Global Average Pooling: $GAP2D(\mathfrak{X})_{\mathfrak{B}}$ represent the global average pooling value for the $\mathfrak{B}_{th}$ channel, is applied to

aggregate spatial information by taking all values in the input tensor $\mathfrak{X}$ is represented as;

$$GAP2D(\mathfrak{X})_{\mathfrak{B}} = 1/n \sum_{i,j} \mathfrak{X}_{i,j} \tag{10}$$

In The above equation, n is the total number of elements in $\mathfrak{X}$.

Batch Normalization: It is subsequently applied to normalize activations, aiming to reduce internal covariates shifts and stabilize training. The BN with input tensor$\mathfrak{X}$, γ and β are learnable scale and shift parameters, μ and σ^2 are the mean and variance, and ϵ is a small constant for numerical stability. The fully connected layer is expressed as:

$$\mathrm{BN}(\mathfrak{X}) = \gamma \cdot \frac{\mathfrak{X} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta, \tag{11}$$

Fully connected layer: The standardized features are subsequently fed into fully connected layers to learn complex patterns of fire and non-fire with data, characterized by:

$$\mathrm{FC}(\mathfrak{X}, \mathcal{W}, \mathfrak{B}) = \sigma\left(\mathcal{W}_{\mathfrak{X}} + \mathfrak{B}\right) \tag{12}$$

where $\mathcal{W}$ represents the weight matrix, $\mathfrak{B}$ is the bias vector, and σ refer the to activation function.

Sigmoid Activation Function: It is commonly applied as the activation function in the output layer of a binary classification model. The features obtained from the fully connected layers are further processed using a sigmoid activation function, converting them into probabilities. The following is the mathematical representation of the sigmoid activation function:

$$\sigma(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{13}$$

Here, Z_i is the $i-th$ element of the input vector. K is the number of classes such as fire and non-fire. The output of the sigmoid function $\sigma(Z)_i$ demonstrates the probability of the presence of fire or non-fire in the input image.

Binary Cross entropy (Log Loss): This loss function is frequently employed for binary classification problems. It calculates the discrepancy between the actual labels and the predicted probabilities for each instance. The binary cross-entropy loss can be as:

$$L(y, \hat{y}) = -\left(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right) \tag{14}$$

Here, y is the true label (0 for non-fire, 1 for fire), and $\hat{y}$ is the predicted probability that the instance belongs to class 1.

Overall, the FGA module improves the localization and classification capability of the network by combining local spatial sensitivity with global semantic context. This design is particularly useful for drone-based fire imagery, where fire regions may be small, partially occluded, irregularly shaped, or embedded in cluttered backgrounds. By using multi-scale convolutions and attention-based weighting, FGA reduces over-reliance on isolated deep feature maps and improves robustness to scale variation, background complexity, and fire-like distractors. The resulting descriptor $\mathbf{F}_{\mathrm{FGA}} \in \mathbb{R}^{C'}$ is passed to the Hierarchical Feature Fusion module for final integration and classification. The qualitative heatmaps and quantitative ablation results further confirm that FGA improves attention to fire-affected regions and reduces false responses from non-fire areas. The complete structure of the module is shown in Fig. 3 of the Supplementary Material.

E. Hierarchical Feature Fusion Mechanism

The Fire-Guided Attention (FGA) module enhances the network's spatial focus on fire regions at multiple resolutions. However, to ensure accurate fire detection, it is crucial to integrate both coarse and fine-scale features into a unified representation. To achieve this, we introduce the Hierarchical Feature Fusion (HFF) mechanism, which consolidates multiscale information from both the FGA and DFF modules. This fusion strengthens the model's discriminative ability by aggregating features from different receptive fields and depths, thereby reducing over-reliance on a single feature map. As illustrated in Fig. 3 of the Supplementary Material, features extracted from EfficientNetB2 are first passed through a 7×7 convolutional layer, denoted as $f_{(x)}$, to capture broad spatial fire patterns. Two additional 1×1 convolutional layers, denoted as $k_{(x)}$ and $l_{(x)}$, are then used to refine channel-wise information. The features from $f_{(x)}$ are further processed through 5×5 and 3×3 convolutional layers, denoted as $g_{(x)}$ and $h_{(x)}$, respectively. These layers help in capturing mid- and fine-grained spatial patterns associated with flickering flame tips or small ignition sources. These layers are critical in distinguishing visually ambiguous fire signals from distractors such as sunlight glare or industrial smoke. The output features are then processed through convolution, down-sampling, and batch normalization to obtain X_1, X_2, and X_3. The features X_1 and X_2 are concatenated to form F_1, while X_3 is concatenated with F_1 to obtain F_2. Subsequently, F_2 is combined with X_4, obtained from $k_{(x)}$, to form F_3. Finally, F_3 is concatenated with X_5, obtained from $l_{(x)}$, and further enriched with features from the original EfficientNetB2 through a 1×1 convolutional layer and global average pooling. This hierarchical fusion process preserves both local pixel-level details and global contextual information, producing a multiscale task-specific descriptor that is robust to variations in altitude, viewpoint, scale, and background clutter in drone-captured imagery.

The feature fusion mechanism is represented as:

$$\mathrm{HFF}_{\oplus} = \left(F_1 \oplus F_2 \oplus \cdots \oplus F_{\text{num\_scales}}\right) \tag{15}$$

where $\oplus$ denotes the concatenation of attention-weighted feature maps from different scales. By integrating features from multiple receptive fields, depths, and attention branches, the HFF module reduces dependence on any single representation and improves adaptability to fire scenes with varying intensity, occlusion, and structural complexity. This is particularly important for UAV-based fire monitoring, where the scale and orientation of fire may vary significantly. Fig. 4 in the Supplementary Material shows that HFF achieves stronger spatial focus than individual DFF and partial FGA configurations. Quantitative results in Section III further confirm that HFF improves performance across the main evaluation metrics, as shown in Tables II, III, IV, and V.

TABLE I
COMPARISON AND ABLATION STUDY OF BASELINE MODELS ON THE DF DATASET. THE BEST NETWORK IS HIGHLIGHTED IN BOLD.

Network	Pre:	F1-Score	Acc.	Sens.	Spec.
EffNetB3	94.25	81.61	78.76	71.96	91.76
DenseNet121	100.0	84.96	82.30	73.86	100.0
InceptionV3	99.12	85.01	82.52	74.42	98.68
EffNetB0	96.02	85.43	83.63	76.95	94.91
MobileNet	77.88	82.82	83.85	88.44	80.24
NASNetMob.	78.76	82.98	83.85	87.68	80.72
MobileNetV2	77.88	82.82	83.85	88.44	80.24
EffNetB1	92.48	85.31	84.07	79.17	90.96
EffNetB2	88.05	**87.09**	**86.95**	86.15	87.78

III. EXPERIMENTAL RESULTS

The following sections describe the dataset, system configuration, implementation details, model evaluation metrics, ablation study, and, finally, the comparison of HFF with SOTA methods. The details of each dataset and a corresponding discussion, System Configuration, and Implementation details are given in the Supplementary Material files.

A. Ablation Study

Ablation studies are employed as a powerful analytical tool to systematically evaluate and dissect the effects of varying components within the chosen models. In the domain of fire classification and pattern recognition, analyzing backbone models and FGA components plays an essential role in the development of fire detection systems. Backbone models, which represent the backbone of the deep neural network architecture, are trained to learn the discriminative features of the input images. The integration of FGA components into the backbone model allows the network to adopt a hierarchical feature selection approach, which enables it to selectively attend to salient regions in the input data, improving fire pattern identification. Our aim is to gain a more comprehensive understanding of these elements and ultimately create more comprehensive models, facilitate better model generalization, and improve overall model performance in real-life applications when it comes to fire classification. Moreover, the analysis may help to highlight the factors that affect the accuracy in fire classification, which will contribute to the development of more effective and reliable fire detection and suppression systems.

1) Backbone Models Analysis: This study covers the full scope of ablation studies, and in particular on backbone model selection. For these analyses, nine different baseline CNN models, including InceptionV3, DenseNet121, MobileNet, MobileNetV2, NASNetMobile, EfficientNetB0, EfficientNetB1, EfficientNetB2, and EfficientNetB3) were used by the DF data for extracting the features and for fire and non-fire classification. It is worth noting that EfficientNetB2 has shown outstanding performance with a commendable accuracy of 86.95% due to its powerful feature extraction capabilities. EfficientNetB3 on the other hand, recorded the least performance with 78.76% on the DF dataset as presented in **Table I**.

TABLE II
COMPARISON AND ABLATION STUDY OF THE FGA COMPONENT ON THE DF DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

FGA Comp.	Pre:	F1-Score	Acc.	Sens.	Spec.
F-X_1X_2	84.07	87.16	87.61	90.48	85.12
F-X_1X_3	85.40	88.53	88.94	91.90	86.36
F-X_2X_3	84.96	87.67	88.05	90.57	85.83
HFF	83.19	**88.47**	**89.16**	94.47	84.98

2) Fire Guided-Attention Component Analysis: The subsequent phase of our investigation involved the utilization of diverse optimizing the Fire Guided-Attention configuration. Our experimental endeavors involved the exploration of diverse combinations of Fire Guided-Attention module to gauge the effectiveness of the HFF attention method across varied compositions. The integration of meticulously optimized attention modules was found to significantly enhance feature extraction, thereby contributing to superior performance in fire classification. Evaluation of different HFF configurations: I) Integration of Fire Guided-Attention with convolutional kernels (7×7) and (5×5) denoted as (F-X_1X_2), II) Integration of Fire Guided-Attention with convolutional kernels (7×7) and (3×3) represented as (F-X_1X_3), III) Integration of Fire Guided-Attention with convolutional kernels (5×5) and (3×3) represented as (F-X_2X_3) and collectively termed as the HFF attention method, these configurations were rigorously examined. Notably, results in **Table II** underscore that the integration of Fire Guided-Attention through the convolutional operation with a kernel size of (7×7) and (5×5) when compared to the other integrated Fire Guided-Attention module, demonstrates comparatively diminished performance. Conversely, the incorporation of Fire Guided-Attention with convolutional kernels (7×7) and (3×3) yields superior results, with accuracy increasing from 87.61% to 88.94%. Particularly noteworthy is the combination of convolutional kernels (5×5) and (3×3) attention modules with the baseline method, showcasing the highest performance at 88.05% accuracy. HFF consistently outperforms regarding accuracy, as detailed in **Table II**. These results underscore the significant role of FGA in maximizing feature representations, minimizing FPR and enhancing classification performance via hierarchical attention weighting. In short, HFF model is the most efficient and effective of all possible mixes, which highlights the effectiveness and efficiency of our suggested model.

IV. PERFORMANCE COMPARISON WITH SOTA

In this subsection, the performance proposed model has been thoroughly investigated in detail and compared its performance with SOTA models using two benchmark datasets as described in the dataset section in detail and our proposed DF dataset. The comparison results presented in Tables III and IV are based on previously reported SOTA methods available for each corresponding dataset. Since some prior studies did not report all evaluation metrics uniformly, only the available performance measures from the respective publications are included to ensure a fair and accurate comparison. Quantitative

TABLE III
COMPARISON OF HFF WITH SOTA MODELS USING FD AND FLAME DATASETS. THE SUPERIOR PERFORMANCE OF HFF IS REFLECTED IN THE RESULTS, WITH THE HIGHEST VALUES HIGHLIGHTED IN **BOLD**.

	Methods	FD Data [32]				FLAME Dataset [33]			
		Acc:	F1-Score	Recall	Pre:	Acc:	F1-Score	Recall	Pre:
TML	FPC [34]	0.53	0.68	0.99	0.52	-	-	-	-
	FD-GCM [35]	0.69	0.74	0.90	0.63	-	-	-	-
	FFD-ANN [36]	0.71	0.72	0.73	0.71	-	-	-	-
Deep learning	ANetFire [8]	87.20	87.90	93.20	83.30	-	-	-	-
	GNetFire [37]	92.30	92.80	98.00	88.00	-	-	-	-
	EMNFire [38]	92.80	93.20	98.70	88.30	-	-	-	-
	EFDNet [30]	95.30	95.40	97.40	93.50	-	-	-	-
	DFAN [14]	96.17	96.00	97.00	96.00	-	-	-	-
	OFAN [39]	96.54	97.00	96.00	97.00	-	-	-	-
	Xception [33]	-	-	-	-	76.23	-	-	-
	Ensemble model [40]	-	-	-	-	85.12	-	-	84.77
	RCFCL [41]	-	-	-	-	88.00	-	-	-
	RDSA [31]	-	-	-	-	93.65	-	-	-
	ACNet [42]	-	-	-	-	97.45	97.12	97.10	98.20
	HFF	**97.52**	97.47	99.56	95.46	**97.50**	97.95	97.91	97.98

TABLE IV
COMPARISON OF HFF WITH SOTA MODELS ON THE DSFD DATASET. THE BEST ACCURACY IS HIGHLIGHTED IN BOLD.

Model	Acc.	F-Score	Recall	Pre:
EFDNet [30]	88.00	87.50	88.00	87.50
DFAN [14]	89.36	89.84	94.00	86.01
ADFireNet [43]	90.86	89.84	90.86	90.90
M-SoftFireNet [44]	93.50	93.51	93.51	93.57
HFF	**96.00**	96.03	95.25	96.83

TABLE V
COMPARISON OF HFF WITH VISION TRANSFORMER-BASED SOTA MODELS ON THE DSFD DATASET. THE HIGHEST SOTA ACCURACY IS IN ITALICS, WHILE HFF IS IN BOLD.

Model	KD Method	MViT-S	MViT-XS	Swin
ViT/32	Soft Target KD	91.33	94.83	–
	DIST	94.17	93.00	–
	OFA-KD	*95.33*	95.00	–
ConvNeXt	Soft Target KD	92.00	92.33	–
	DIST	95.00	95.00	–
	OFA-KD	95.17	*95.50*	–
Swim-Transformer	–	–	–	*93.67*
HFF	–	–	–	**96.00**

and qualitative analysis of the HFF and the SOTA are outlined in detail in the subsequent section.

A. Quantitative Results

To evaluate the effectiveness of the proposed HFF framework, we compared its performance with traditional machine learning (TML), deep learning (DL), and vision transformer-based methods on the proposed DF dataset and three benchmark datasets: FD, FLAME, and DSFD. The results demonstrate the suitability of HFF for robust fire scene classification across diverse fire, non-fire, and fire-like visual conditions.

1) HFF Comparison with TML Approaches: The proposed HFF framework was first compared with TML-based fire detection methods on the FD dataset using the evaluation protocol reported in [39]. As shown in **Table III**, HFF outperforms existing TML methods across the main evaluation metrics, including accuracy, recall, F1-score, and precision. In particular, HFF achieves an accuracy of 97.52% on the FD dataset, indicating a clear improvement over handcrafted feature-based methods. This performance gain shows that the proposed framework can learn complex fire-related representations directly from visual data, reducing the dependence on manually designed features.

2) HFF Comparison with DL Approaches: The HFF framework was further evaluated against recent DL-based fire detection methods on FD, FLAME, DSFD, and the proposed DF dataset. The results confirm that combining fire-specific feature extraction with hierarchical feature fusion improves classification performance across different datasets and sensing conditions. As reported in **Table III**, HFF achieves the best overall performance on the FD dataset, with 97.52% accuracy, 97.47% F1-score, 99.56% recall, and 95.46% precision. Among the compared methods, ANetFire [8] records the lowest performance, while OFAN [39] and DFAN [14] are the closest competitors in terms of accuracy. Although OFAN and DFAN report higher precision values of 97% and 96%, respectively, HFF achieves better overall balance across recall, F1-score, and accuracy. The strong performance achieved by HFF on the FD dataset demonstrates that the proposed hierarchical feature fusion and fire-guided attention mechanisms generalize effectively across heterogeneous visual perspectives. On the FLAME dataset, Xception [33] obtains the lowest accuracy of 76.23%, while ACNet [42] achieves 97.45%. The proposed HFF framework further improves the result, reaching 97.50% accuracy, with 97.95% F1-score, 97.91% recall, and 97.98% precision. On the proposed DF dataset, HFF also demonstrates strong classification ability, achieving 97.50% accuracy, 97.95% F1-score, 97.91% recall, and 97.98% precision. The comparison on the DSFD dataset is presented in **Table IV**. Among the evaluated methods, EFDNet [30] reports the lowest accuracy of 88.00%, whereas the proposed HFF model achieves the highest accuracy of 96.00%, representing an 8.00% improvement. DFAN [14] and ADFireNet [43] achieve accuracies of 89.36% and 90.86%, respectively, while M-SoftFireNet [44] reaches 93.50%. However, HFF remains superior by a margin of 2.50% over M-

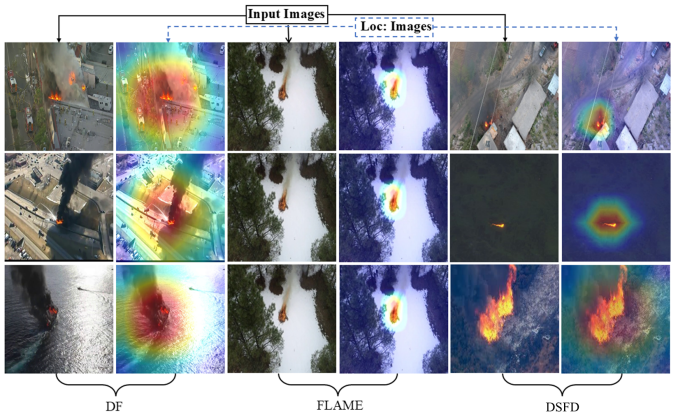

Fig. 2. Visual results obtained by the proposed model across three distinct datasets (DF, FLAME, and DSFD) using the HFF model. The first, third, and fifth columns show input images from the DF, FLAME, and DSFD datasets, respectively, while the second, fourth, and sixth columns display the visual localization results achieved by the HFF model on the same datasets.

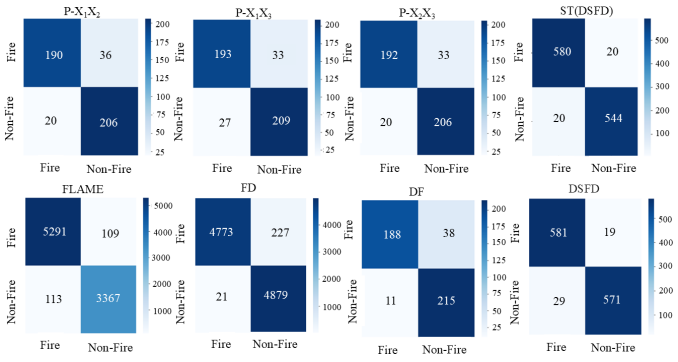

Fig. 3. Comparison of confusion matrices: ablation study vs. proposed method. The first row displays the confusion matrix for three ablation study variants, which include FX_1X_2, FX_1X_3, and FX_2X_3, alongside the Swim Transformer (ST) which is evaluated on the DSFD dataset. The second row shows the confusion matrix of the HFF framework across four distinct datasets: Flame, FD, DF and ASDF. The results illustrate the HFF model's superior accuracy and generalization in distinguishing fire and non-fire instances under diverse environmental conditions.

SoftFireNet. A similar trend is observed for the F1-score, where HFF obtains the highest value of 96.03%, compared with 87.50% for EFDNet, 89.84% for DFAN and ADFireNet [43], and 93.51% for M-SoftFireNet [44]. HFF also achieves the highest recall and precision values of 95.25% and 96.83%, respectively. Overall, these results show that the proposed HFF framework consistently outperforms existing TML and DL-based methods. The improvement can be attributed to the joint contribution of DFF, FGA, and hierarchical feature fusion, which enables the model to capture fire patterns at multiple scales while reducing confusion with fire-like background regions.

3) HFF Comparison with Vision Transformer Approaches: To further assess the generalization capability of HFF, we compared it with recent vision transformer-based models on the DSFD dataset. Specifically, we considered MobileViT-S and MobileViT-XS models trained with different knowledge distillation (KD) strategies, including soft target KD, stronger teacher distillation (DIST), and one-for-all KD (OFA-KD), as reported in [20]. We also implemented the Swin Transformer for additional comparison. As summarized in **Table V**, the best MobileViT-XS result is 95.50%, obtained using OFA-KD with ConvNeXt as the teacher model. Under the same setting, MobileViT-S achieves 95.33%. In our experiments, the Swin Transformer obtains an accuracy of 93.67% on the DSFD dataset. In comparison, the proposed HFF framework achieves the highest accuracy of 96.00%. These findings indicate that HFF can outperform transformer-based alternatives on the DSFD dataset while avoiding the heavy pretraining commonly required by transformer architectures. Therefore, the proposed framework provides a favorable balance between accuracy and practical efficiency, making it suitable for UAV-based fire detection and disaster monitoring applications.

B. Qualitative Results

The qualitative performance of the proposed HFF framework was examined on four datasets, namely FD, FLAME, DF, and DSFD, using Grad-CAM visualizations, confusion matrices, and ROC and Precision–Recall (PR) curves. These analyses complement the quantitative results by showing not only how accurately the model classifies fire and non-fire scenes, but also how it reaches those decisions under different sensing and environmental conditions.

Grad-CAM was applied to visualize the regions of the image that had the most significant contribution to the model predictions. As illustrated in Fig. 2. The proposed HFF framework is always focused on useful fire features and inhibits the irrelevant background reactions. This fact can be attributed to the advantage of integrating both hierarchical feature fusion and the proposed Fire Guided-Attention (FGA) module, which assists the network to retain not only the local fire properties but also the greater contextual information. The visual performance shows that the model has the ability to detect fire-related areas in images with varying drone-view characteristics and reduce responses to non-fire scenarios. Further examples in Fig. 6 in the supplementary file further highlight that the model can still be used in difficult situations such as heavy smoke, fog, night scenes and daytime variations. In the cases where the visibility is poor, the activation maps are still localized around the true fire areas, implying that the model is acquiring pertinent fire structure, and not making use of incidental cues in the dataset. When comparing with DFAN [14] in Fig. 7 of the supplementary file, it is clear that the suggested HFF model generates smaller and spatially accurate activations, and the DFAN [14] tends to focus on extraneous or overly general areas.

The confusion matrices depicted in Fig. 8 in the supplementary file and Fig. 3 gives further insight into class-wise prediction patterns over various settings of backbone and various datasets. These findings demonstrate that the proposed framework provides a balanced classification of fire and non-fire samples and is stable in various feature extraction setups. The ablation study also confirms that the enhancement of the model of distinguishing between similar fire scenes and non-fire scenes is enhanced by the integration of the FGA module. The consistency of the learned representation is also supported

by representative predictions in Fig. 9 in the supplementary file, which show that the diverse test samples are correctly labeled by the proposed model.

To evaluate the discriminative power of the suggested framework further, ROC and PR curves were created using each of the DF, FLAME, FD and DSFD datasets, which can be seen in Fig. 10 of the supporting file. As every dataset represents a different sensing situation, independent testing gives a better understanding of how the model generalizes. The HFF framework offers the highest scores to the AUC and AP of 0.99 on the DF dataset, which points to excellent performance on images captured by drones with a high level of geographic and environmental diversity. The model gave an AUC of 1.00 and an AP of 1.00 on the FLAME and FD datasets, indicating that it can discriminate very effectively with both UAV-based and ground-based fire scenes. The model achieved an AUC of 0.99 and an AP of 0.98 on the DSFD dataset, which comprises drone and satellite imagery in different atmospheric conditions. Overall, these results confirm the idea that the specified HFF framework can be used to provide the appropriate localization, uniform classification, and high generalization in diverse fire monitoring scenarios.

V. Discussion

The proposed HFF model also performs better in UAV-based fire detection, which is an empirically tested model with various datasets including DF, FD, FLAME and DSFD. The HFF model as shown in Tables III to V, is always superior to the conventional CNN-based and attention-based models in accuracy of detection and other performance indicators. Furthermore, Figure 2 to 3 and also the visual results are provided in the supplementary files, displaying the Qualitative results with Grad-CAM visualization and heat maps, which additionally reveal the capability of the model to efficiently localize the fire locations even under unfavorable conditions, such as fog, occlusion and low visibility. All these results validate the strength and versatility of HFF in various fire detection conditions. The architecture of HFF is its strength as it is theoretically motivated and technically based and was created to solve the current challenges of the current fire detection systems. In contrast to earlier techniques, which are based heavily on sequential layer feature maps, which have problems with scale variation, HFF presents a more modular structure, with each element architecture designed to overcome a particular difficulty of fire detection.

The Discriminative Fire Feature (DFF) extractor is built on EfficientNetB2 which was selected as the foundation of the HFF framework.This selection was made by performing extensive empirical comparisons of the nine CNN backbones models, as demonstrated in Table I. EfficientNetB2 proved better performance its compound scaling algorithm that also enables depth, width and resolution optimization, which is essential in capturing fire features with high fidelity in UAV imagery. In theory, its design provides both accuracy and computational efficiency, which is particularly applicable to UAV-based applications. In theory, its design provides both accuracy and computational efficiency, which is particularly applicable in mobile aerial applications. In technical terms, its use of squeeze-and-excitation blocks and swish activation also improves its capacity to extract semantically rich and contextually relevant features of visually complex scenes.

To overcome the drawback of dealing with the scale variation which is one of the main problems of fire detection models, the FGA module has been proposed. This module uses multi-kernel convolutions (3x3, 5x5, 7x7) to replicate the receptive field size. FGA is theoretically based on multiscale representation learning and enhances the model sensitivity to both fine-grained and broad contextual features. It is particularly effective in scenarios when dealing with small or partially blocked fires, as demonstrated in Fig.6 and 7 in the supplementary file, which show the model is localized and focused on the important fire areas.

The hierarchical feature fusion mechanism combines the feature maps of the DFF and FGA modules with a multi-level and systematic concatenation method. Instead of using a single feature map, the design will pool spatially localized information with high-level semantic context, enabling the model to better make sense of complex visual scenes. The fusion strategy boosts contextual knowledge and increases detection accuracy by aligning features across scales. As a result, it minimizes false positives and enhances robustness especially in environments where false positives could have been caused by fire-like patterns or occlusions.

The HFF framework, including DFF extractor, which performs robust base feature extraction, FGA module, which refines the features in multi-scale, and the hierarchical feature fusion mechanism, which integrates the features in context, is specifically created to address the fundamental issues in fire detection. These may be classified as loss of spatial resolution, sensitivity to scale changes, and dependence upon feature representation with few features. The modules are specifically designed to address these issues associated with fires and theory and empirical evidence are used to support their effectiveness. The HFF model provides a generalised, interpretable and powerful solution by matching innovations with domain needs for detection. It is a significant step in the development of UAV-based fire detection, and the potential of consequences on disaster management or environmental surveillance is considerable.

VI. Computational Complexity Analysis

Real-time processing is essential for UAV-based fire detection, where timely response is crucial. The performance of the proposed Hierarchical Feature Fusion (HFF) model was quantitatively evaluated in terms of the model size, number of parameters, and inference speed for both the GPU and CPU platforms. These metrics are summarized and compared with the existing state-of-the-art fire detection models in Table VI.

The HFF model was compared against state-of-the-art fire detection approaches including ANetFire [8], ResNetFire [45], EMNFire [38], GNetFire [37], DFAN [14], SE-EFFNet [46], ViT-B-32 [47], FlareNet [48], ADFireNet [43], and MAFireNet [49], as summarized in Table VI. ANetFire [8], with the largest model size of 233 MB and 60 million parameters,

achieved only 17 FPS on CPU, highlighting the trade-off between size and efficiency. ResNetFire [45] and EMNFire [38], while smaller, suffered from limited FPS on CPU (2.4 and 6.3, respectively) despite competitive accuracy. SE-EFFNet [46] and DFAN [14] demonstrated promising GPU performance but had larger model sizes or slower CPU inference. The HFF framework has a model size of 54.42 MB and 14.27 million parameters, achieving 80.93 FPS on GPU and 25.51 FPS on CPU. Compared to ANetFire [8], this corresponds to a reduction of approximately 76.6% in size and 76.2% in parameters, while substantially increasing inference speed. Compared to MAFire-Net [49], a recent method with 74.43 MB and 22.6 million parameters, HFF attains higher FPS on both GPU (80.93 vs 78.31) and CPU (25.51 vs 14.32), indicating more efficient utilization of computational resources.

The improved efficiency of HFF is largely due to the hierarchical feature fusion mechanism and multi-scale attention operations in the FGA module, which selectively emphasize fire-relevant features without introducing excessive computational overhead. These design choices ensure that HFF can operate effectively on UAV platforms with limited processing power, providing a practical balance between model complexity, inference speed, and detection accuracy.Overall, the analysis demonstrates that HFF outperforms prior models in terms of computational efficiency and real-time applicability, making it a suitable candidate for deployment in real-world UAV-based fire monitoring systems.

TABLE VI
COMPARISON OF MODEL SIZE, PARAMETERS, AND INFERENCE SPEED WITH EXISTING FIRE DETECTION MODELS.

Reference	Size (MB)	Parameters (Millions)	Inference Time	
			GPU FPS	CPU FPS
ANetFire [8]	233.0	60.0	–	17
ResNetFire [45]	98.0	25.6	57.3	2.4
EMNFire [38]	13.2	3.5	34.0	6.3
GNetFire [37]	43.30	–	48.2	4.3
DFAN [14]	83.63	23.9	70.55	12.90
SE-EFFNet [46]	47.75	12.4	45.0	8.0
ViT-B-32 [47]	89.82	88.22	67.0	20.0
FlareNet [48]	49.1	–	75.0	14.0
ADFireNet [43]	38.0	7.2	72.5	22.0
MAFire-Net [49]	74.43	22.6	78.31	14.32
HFF model	**54.42**	**14.27**	**80.93**	**25.51**

VII. CONCLUSION

This paper presents a deep learning model that uses a hierarchical approach to learning features and a Fire Guided-Attention system to enhance discriminative features representation in fire detection. Ablation experiment of the challenging DF dataset validates the efficiency and ability of the proposed design to perform in complex aerial imaging. We also compared HFF with current state-of-the-art approaches on the FLAME, FD and DSFD data, where HFF has demonstrated highly effective performance on a variety of evaluation measures, demonstrating its usefulness in a wide variety of UAV-based fire monitoring tasks, such as building fire, forest fire and complex outdoor environment. The model also remains effective in both clear and challenging environments. Its performance proves to be reliable on small, distant, and partially obscured fire areas with varying viewpoints. These features render it suitable in real time detection of fires in time sensitive scenarios. **Limitations**: The present research can only be applied to fire detection and classification in complicated surveillance and in air monitoring systems. The proposed framework is not aimed at fire mapping, estimation of fire intensity, and small-scale classification of various types of fire. **Future Direction;** The future will be done to multi-modal fire analysis by using thermal infrared, multispectral and visual images. We will also consider vision language models to enhance adaptability of the model and increase the applicability of the model in real world fire monitoring activities.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY

A sample subset of the constructed DF dataset is available for review at the following link: https://drive.google.com/file/d/1dLxg6DSz2jAmGYiOkRCWcAX14zT8XbBE/view?usp=sharing. The full DF dataset will be made publicly available after publication of the paper.

REFERENCES

[1] L. A. F. D. Margaret Stewart, "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/;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: 2025-01-23.

[2] F. Yuan, G. Wang, Q. Huang, and X. Li, "A newton interpolation network for smoke semantic segmentation," *Pattern Recognition*, vol. 159, p. 111119, 2025.

[3] S. Danish, M. J. Piran, S. U. Khan, M. A. Khan, L. M. Dang, Y. Zweiri, H.-K. Song, and H. Moon, "Vision-based fire management system using autonomous unmanned aerial vehicles: a comprehensive survey," *Artificial Intelligence Review*, vol. 59, no. 1, p. 16, 2025.

[4] X. Yang, Z. Hua, L. Zhang, X. Fan, F. Zhang, Q. Ye, and L. Fu, "Preferred vector machine for forest fire detection," *Pattern Recognition*, vol. 143, p. 109722, 2023.

[5] X. Chen, Q. An, and K. Yu, "Fire identification based on improved multi feature fusion of ycbcr and regional growth," *Expert Systems with Applications*, vol. 241, p. 122661, 2024.

[6] X. Gong, J. Wang, Q. Ren, K. Zhang, E.-S. M. El-Alfy, and J. Mańdziuk, "Embedded feature selection approach based on tsk fuzzy system with sparse rule base for high-dimensional classification problems," *Knowledge-Based Systems*, vol. 295, p. 111809, 2024.

[7] D. Y. Chino, L. P. Avalhais, J. F. Rodrigues, and A. J. Traina, "Bowfire: detection of fire in still images by integrating pixel color and texture analysis," in *2015 28th SIBGRAPI conference on graphics, patterns and images*. IEEE, 2015, pp. 95–102.

[8] Y. H. Habiboğlu, O. Günay, and A. E. Çetin, "Covariance matrix-based fire and flame detection method in video," *Machine Vision and Applications*, vol. 23, pp. 1103–1113, 2012.

[9] V. K. Singh, S. Chakraborty, R. Singh, R. S. Rathore, and W. Jiang, "Bias-aware data quality enhancement for forest fire detection in ai-based remote sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[10] H. Tao, Q. Duan, M. Lu, and Z. Hu, "Learning discriminative feature representation with pixel-level supervision for forest smoke recognition," *Pattern Recognition*, vol. 143, p. 109761, 2023.

[11] H. Li, Z. Ma, S.-H. Xiong, Q. Sun, and Z.-S. Chen, "Image-based fire detection using an attention mechanism and pruned dense network transfer learning," *Information Sciences*, vol. 670, p. 120633, 2024.

[12] I. Ziadi, N. Essaddi, and M. Besbes, "Ai-driven classification of tsunami-generating earthquakes: Harnessing random forest, svm, and logistic regression for early detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[13] S. U. Khan, S. Danish, E. Iqbal, D. Gupta, Y. Zweiri, and Y. Abdulrahman, "Firemamba: An efficient visual representation learning framework for fire detection," *Alexandria Engineering Journal*, vol. 132, pp. 133–144, 2025.

[14] H. Yar, T. Hussain, M. Agarwal, Z. A. Khan, S. K. Gupta, and S. W. Baik, "Optimized dual fire attention network and medium-scale fire classification benchmark," *IEEE Transactions on Image Processing*, vol. 31, pp. 6331–6343, 2022.

[15] H. Liz-López, J. Huertas-Tato, J. Pérez-Aracil, C. Casanova-Mateo, J. Sanz-Justo, and D. Camacho, "Spain on fire: A novel wildfire risk assessment model based on image satellite processing and atmospheric information," *Knowledge-Based Systems*, vol. 283, p. 111198, 2024.

[16] H. Yar, Z. A. Khan, T. Hussain, and S. W. Baik, "A modified vision transformer architecture with scratch learning capabilities for effective fire detection," *Expert Systems with Applications*, vol. 252, p. 123935, 2024.

[17] K. Mardani, N. Vretos, and P. Daras, "Transformer-based fire detection in videos," *Sensors*, vol. 23, no. 6, p. 3035, 2023.

[18] Z. Wang, Z. Wang, H. Zhang, and X. Guo, "A novel fire detection approach based on cnn-svm using tensorflow," in *Intelligent Computing Methodologies: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part III 13*. Springer, 2017, pp. 682–693.

[19] F. Saeed, A. Paul, P. Karthigaikumar, and A. Nayyar, "Convolutional neural network based early fire detection," *Multimedia Tools and Applications*, vol. 79, no. 13, pp. 9083–9099, 2020.

[20] S. Jangirova, B. Jankovic, W. Ullah, L. U. Khan, and M. Guizani, "Real-time aerial fire detection on resource-constrained devices using knowledge distillation," *arXiv preprint arXiv:2502.20979*, 2025.

[21] K. Niu, C. Wang, J. Xu, J. Liang, X. Zhou, K. Wen, M. Lu, and C. Yang, "Early forest fire detection with uav image fusion: A novel deep learning method using visible and infrared sensors," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[22] P. Bu, R. W. Aslam, A. Quddoos, N. Y. Rebouh, M. N. Ahmad, R. M. Zulqarnain, Q. Abbas, and Y. Said, "Multi-sensor data fusion for quantifying agricultural fire impacts on air quality and environmental degradation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[23] J. Zhang, G. Zhang, Y. Kang, Y. Dong, Y. Liu, S. Xie, and H. Xu, "Determination of forest fire intensity level using multi-temporal satellite remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[24] Y. Li, J. Liu, E. E. Maeda, X. Li, P. Pellikka, and J. Heiskanen, "Detection of forest burned area using a spatiotemporal algorithm based on spectral index time series data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 28 971–28 985, 2025.

[25] M. I. Keskes and M. D. Niţă, "Digital twins in forest management using scalable deep learning pipeline," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[26] R. A. Loehman, M. M. Friggens, C. I. Constan, and A. Steffen, "Relationship of satellite-derived fire severity to archaeological fire severity and fire effects in the jemez mountains, new mexico, usa," *Fire Ecology*, 2026.

[27] A. C. Muthiuru, J. D. Millington, K. Chan, and E. J. Tebbs, "Burned area trends and fire susceptibility in protected areas of kenya: the potential rolesof human activities and climate," *Fire Ecology*, 2026.

[28] X. Liu, L. Sun, Y. Fan, C. Wang, and H. Yu, "A multiple spectral criteria active fire detection algorithm supported by hyperspectral dataset," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[29] S. Majid, F. Alenezi, S. Masood, M. Ahmad, E. S. Gündüz, and K. Polat, "Attention based cnn model for fire detection and localization in real-world images," *Expert Systems with Applications*, vol. 189, p. 116114, 2022.

[30] S. Li, Q. Yan, and P. Liu, "An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism," *IEEE Transactions on Image Processing*, vol. 29, pp. 8467–8475, 2020.

[31] Z. Guan, X. Miao, Y. Mu, Q. Sun, Q. Ye, and D. Gao, "Forest fire segmentation from aerial imagery data using an improved instance segmentation model," *Remote Sensing*, vol. 14, no. 13, p. 3159, 2022.

[32] P. Foggia, A. Saggese, and M. Vento, "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*, vol. 25, no. 9, pp. 1545–1556, 2015.

[33] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial imagery pile burn detection using deep learning: The flame dataset," *Computer Networks*, vol. 193, p. 108001, 2021.

[34] T. Celik, H. Ozkaramanli, and H. Demirel, "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. IEEE, 2007, pp. I–1205.

[35] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," *Fire safety journal*, vol. 44, no. 2, pp. 147–158, 2009.

[36] D. Zhang, S. Han, J. Zhao, Z. Zhang, C. Qu, Y. Ke, and X. Chen, "Image based forest fire detection using dynamic characteristics with artificial neural networks," in *2009 international joint conference on artificial intelligence*. IEEE, 2009, pp. 290–293.

[37] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, "Convolutional neural networks based fire detection in surveillance videos," *Ieee Access*, vol. 6, pp. 18 174–18 183, 2018.

[38] K. Muhammad, S. Khan, M. Elhoseny, S. H. Ahmed, and S. W. Baik, "Efficient fire detection for uncertain surveillance environment," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3113–3122, 2019.

[39] N. Dilshad, S. U. Khan, N. S. Alghamdi, T. Taleb, and J. Song, "Towards efficient fire detection in iot environment: a modified attention network and large-scale dataset," *IEEE Internet of Things Journal*, 2023.

[40] R. Ghali, M. A. Akhloufi, and W. S. Mseddi, "Deep learning and transformer approaches for uav-based wildfire detection and segmentation," *Sensors*, vol. 22, no. 5, p. 1977, 2022.

[41] S. Treneska and B. R. Stojkoska, "Wildfire detection from uav collected images using transfer learning," in *Proceedings of the 18th International Conference on Informatics and Information Technologies, Skopje, North Macedonia*, 2021, pp. 6–7.

[42] A. M. Islam, F. B. Masud, M. R. Ahmed, A. I. Jafar, J. R. Ullah, S. Islam, S. Shatabda, and A. M. Islam, "An attention-guided deep-learning-based network with bayesian optimization for forest fire classification and localization," *Forests*, vol. 14, no. 10, p. 2080, 2023.

[43] H. Yar, W. Ullah, Z. A. Khan, and S. W. Baik, "An effective attention-based cnn model for fire detection in adverse weather conditions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 206, pp. 335–346, 2023.

[44] H. Yar, Z. A. Khan, I. Rida, W. Ullah, M. J. Kim, and S. W. Baik, "An efficient deep learning architecture for effective fire detection in smart surveillance," *Image and Vision Computing*, vol. 145, p. 104989, 2024.

[45] J. Sharma, O.-C. Granmo, M. Goodwin, and J. T. Fidje, "Deep convolutional neural networks for fire detection in images," in *Engineering Applications of Neural Networks: 18th International Conference, EANN 2017, Athens, Greece, August 25–27, 2017, Proceedings*. Springer, 2017, pp. 183–193.

[46] Z. A. Khan, T. Hussain, F. U. M. Ullah, S. K. Gupta, M. Y. Lee, and S. W. Baik, "Randomly initialized cnn with densely connected stacked autoencoder for efficient fire detection," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105403, 2022.

[47] M. Shahid and K.-l. Hua, "Fire detection using transformer network," in *Proceedings of the 2021 international conference on multimedia retrieval*, 2021, pp. 627–630.

[48] B. Yousaf, A. F. Mirza, M. Irfan, M. Mansoor, and Z. Yang, "Flarenet: A feature fusion based method for fire detection under diverse conditions," 2024.

[49] T. Khan, Z. A. Khan, and C. Choi, "Enhancing real-time fire detection: An effective multi-attention network and a fire benchmark," *Neural Computing and Applications*, vol. 37, no. 18, pp. 11 693–11 707, 2025.