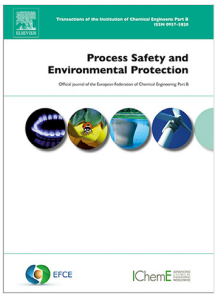
Journal Pre-proof



Masked autoencoder-based vision framework for robust fire detection in complex environments

Hanxiang Wang, Muhammad Fayaz, Awais Ahmad, Yanfen Li, Tan N. Nguyen, L. Minh Dang

PII: S0957-5820(25)01286-8
DOI: https://doi.org/10.1016/j.psep.2025.108019
Reference: PSEP 108019

To appear in: *Process Safety and Environmental Protection*

Received date : 13 July 2025
Revised date : 8 October 2025
Accepted date : 13 October 2025

Please cite this article as: H. Wang, M. Fayaz, A. Ahmad et al., Masked autoencoder-based vision framework for robust fire detection in complex environments. *Process Safety and Environmental Protection* (2025), doi: https://doi.org/10.1016/j.psep.2025.108019.

Masked Autoencoder-Based Vision Framework for Robust Fire Detection in Complex Environments

Abstract

Vision-based fire detection has become an increasingly important focus in computer vision, driven by the growing need for early warning systems and public safety in surveillance environments. While conventional models have primarily relied on color-based features to distinguish fire from background, maintaining high detection accuracy while ensuring computational efficiency remains a persistent challenge, particularly in real-time surveillance systems. To address this, we introduce a novel fire detection framework grounded in masked autoencoding and Vision Transformers (ViT), designed to balance detection performance with scalable deployment. Our architecture leverages self-supervised learning to reconstruct masked visual regions, enhancing the encoder's ability to capture fine-grained fire cues in complex scenarios. The integration of global attention and hierarchical context modeling enables the system to distinguish between fire and visually similar non-fire patterns, such as reflections and artificial lighting, under diverse environmental conditions. Unlike prior models that are sensitive to background noise or rely heavily on channel saliency, our approach learns robust representations through reconstruction objectives, eliminating the need for hand-crafted modules. Extensive experiments conducted on five benchmark datasets: BWF, DQFF, LSFD, DSFD, FG and DFAN demonstrate consistent improvements over existing methods, with notable gains of 2.5% on BWF, 2.2% on DQFF, 1.42% on LSFD, 1.8% on DSFD, 1.14% on FG and 1.10% on DFAN. The proposed model also maintains computational efficiency and generalizes effectively across a wide range of fire conditions, supporting its deployment in practical, real-time systems.

Keywords: Fire Detection, Masked Autoencoder, Vision Transformer, Self-Supervised Learning, Feature Reconstruction, Attention Mechanism.

1. Introduction

Fires represent a severe threat to human life and property due to their rapid and often uncontrollable spread, especially in densely populated regions such as urban residential areas, transportation hubs, and forested environments [1]. Ensuring prompt detection of fire outbreaks is crucial to mitigate damage and improve public safety in various domains, including residential, commercial, and industrial settings.

Conventional fire detection (FD) systems typically employ environmental sensors like smoke, temperature, and particle detectors [2]. These sensors are cost-effective and relatively easy to deploy, particularly in confined indoor environments. However, their effectiveness diminishes significantly in open or large-scale outdoor scenarios. Moreover, these systems often activate only when they directly detect fire by-products such as heat or smoke, potentially delaying the response time and reducing the chances of early intervention [3, 4].

To overcome these limitations, there has been a growing interest in vision-based FD technologies which utilize cameras as sensors to monitor large areas in real-time [5, 6]. These systems offer broader coverage, quicker response, and better adaptability to various environmental conditions[7]. Consequently, numerous vision-based methods have been proposed, mainly categorized as traditional machine learning (TML) and deep learning (DL) approaches[8].

TML-based approaches are based on traditional feature extraction techniques, such as flame texture, color, and motion patterns [6]. The success of these methods depends heavily on the quality and relevance of the features designed manually. However, it is challenging to design a robust global feature extraction due to the diverse and dynamic nature of fire. Variations in flame color caused by different combustible materials, lighting conditions, and environmental influences such as wind or temperature fluctuations contribute to the unpredictability of flame behavior. These factors can hinder consistent FD and increase the likelihood of false positives. To effectively harness TML-based methods, it is essential to master the challenge of attaining a high true positive rate while minimizing the false alarm rate.

In response to these challenges, deep learning has emerged as a powerful alternative, offering end-to-end learning capabilities and superior performance across a variety of computer vision tasks [9, 10]. DL-based models automatically learn discriminative features from large datasets, enabling them to generalize effectively to new and unseen fire scenarios. This capacity to extract complex patterns has led to substantial improvements in detection accuracy and robustness. Notably, several DL-based techniques have demonstrated enhanced reliability over TML approaches, especially under variable and uncertain conditions [11].

Despite these advancements, DL-based FD systems are not without limitations. Their performance can degrade in visually complex scenes, such as when fire-colored objects are present or when the fire source is distant and small in the frame, as il-

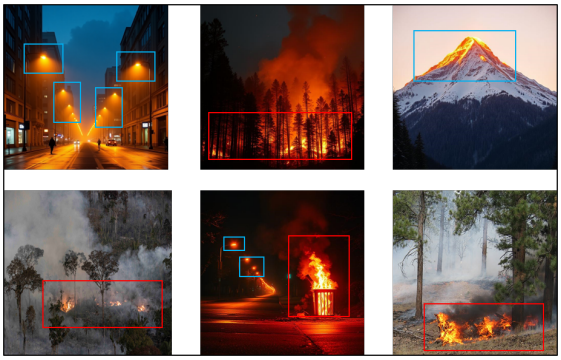


Figure 1: Sample images illustrating challenging scenarios for FD. The red bounding boxes indicate actual fire regions, while the blue bounding boxes highlight visually similar non-fire regions such as lighting artifacts, reflections, or sunlight patches. These examples demonstrate the difficulty in distinguishing real fires from fire-like patterns, particularly under varying environmental and lighting conditions.

lustrated in Figure . These challenges highlight ongoing gaps in the literature and motivate the development of improved detection techniques, which are addressed in the proposed methodology section.

1.1. Research Gap

Despite notable progress in FD technologies, significant challenges remain unresolved, particularly in achieving timely and reliable detection in complex and dynamic environments. Traditional sensor-based systems, while cost-effective and simple to deploy, are constrained by their limited range and delayed responsiveness, especially in outdoor or large-scale settings. These systems often rely on direct detection of smoke or heat, which restricts their ability to identify fires in their early stages. On the other hand, vision-based approaches, although more promising in terms of coverage and responsiveness, also encounter limitations.

Traditional Machine Learning (TML) techniques depend heavily on manually crafted features such as flame color, texture, and motion, which are highly susceptible to environmental variations including lighting conditions, background clutter, and fire appearance diversity. This dependence leads to inconsistent detection performance and a high rate of false alarms.

While Deep Learning (DL)-based models have demonstrated superior accuracy and generalization in FD tasks, their effectiveness declines in certain complex scenarios. Specifically, DL models struggle with false detections when visually similar objects (e.g., fire-colored materials) are present, or when fire is located at long distances, making it less visible in the frame. Additionally, many existing DL methods have not been rigorously evaluated across a wide range of real-world conditions, limiting their practical applicability.

Therefore, there is a clear need for an advanced FD framework that combines the strengths of deep learning with robust feature representation, capable of handling real-time detection in diverse and challenging environments while minimizing false positives and improving detection accuracy across varying scales and contexts.

1.2. Main Contributions

To address the challenges and research gaps in vision-based FD, this study introduces a novel FD framework based on masked autoencoding and transformer architectures. The proposed method aims to improve early FD accuracy and robustness in complex environments by leveraging advanced feature learning and contextual reasoning. The main contributions of this work are as follows:

- We propose an Image Masked Autoencoder (ImageMAE) based FD framework that efficiently learns rich and discriminative fire-related features through a self-supervised reconstruction task. The framework uses an asymmetric encoder-decoder design, where the encoder processes only visible image patches, significantly reducing computational overhead.
- A Vision Transformer (ViT)-based encoder is utilized for feature extraction, capturing long-range dependencies and complex fire patterns such as varying flame shapes, colors, and textures. This enhances the model's ability to distinguish between real fires and fire-like objects in challenging visual conditions.
- The reconstruction module incorporates a novel pixel-level reconstruction loss with optional normalized pixel targets, improving feature invariance to environmental factors such as lighting changes and smoke occlusion. This leads to more robust representations suitable for early-stage FD.
- We conduct comprehensive experiments on a diverse fire dataset, demonstrating that the proposed method achieves superior FD accuracy compared to baseline approaches. Additionally, the model's lightweight decoder and efficient masking strategy enable scalability and real-time applicability in surveillance systems.

1.3. Study Outline

The remainder of this paper is organized as follows: Section II reviews recent literature on FD methods, highlighting existing challenges and advances. Section III details the proposed ImageMAE-based FD framework, including architecture and training strategies. Section IV presents the datasets used, experimental results, and a comprehensive analysis of the model's performance. The conclusion and future direction are presented in Section V.

2. Related Work

The development of early and accurate FD technology using vision sensors has been an active research area and can be broadly categorized into Traditional Machine Learning (TML) methods and Deep Learning (DL) methods. This section provides a detailed overview of both approaches.

2.1. Traditional Machine Learning Methods

Traditional machine learning (TML) approaches primarily focus on handcrafted features based on fire characteristics such as color, shape, motion, and texture [12]. Early color-based FD methods were proposed by [13, 14]; however, these methods suffered from high false positive rates and limited accuracy, restricting their practical application. [15] introduced an auto-adaptive edge detection technique to identify fire regions, while [16] applied methods like K-means clustering, optical flow, logistic regression, and temporal smoothing for FD. [17] employed multiple classification algorithms and combined their outputs for accurate real-time fire scene classification. [18] used covariance features with Support Vector Machines (SVM) for fire scene classification. Teng et al. [19] applied hidden Markov models to detect moving fire pixels extracted from pixel clusters. [20] developed BoWFire, a novel approach that fused color features with super-pixel texture discrimination to improve FD. While traditional machine learning methods have contributed to fire detection, they rely heavily on handcrafted features and standard classifiers, which are error-prone and labor-intensive. They often confuse fire with fire-like objects and perform poorly under varying lighting, smoke, or occlusion. These challenges highlight the need for approaches that can automatically learn robust features. Our ImageMAE-ViT framework addresses this by combining self-supervised learning with global context modeling, enabling accurate and efficient fire detection across diverse conditions.

2.2. Deep Learning Methods

Deep learning (DL) techniques automatically learn features from raw image data and have demonstrated superior performance over traditional methods in FD. Sharma et al. (2017) introduced a challenging fire dataset and evaluated models such as ResNet50 [21] and VGG16 [22], with ResNet50 achieving promising results. [23] designed a convolutional neural network (CNN) for fire scene classification. [24] combined local binary patterns with AdaBoost to isolate fire regions, which were then input to a CNN for feature extraction and classification. [25] utilized pretrained GoogLeNet weights to balance efficiency and accuracy for FD. In another study, they fine-tuned AlexNet for indoor and outdoor fire scenarios, employing adaptive camera prioritization strategies. [26] used a modified InceptionV1 architecture for efficient FD. [27] compared various classifiers, including MLP, AdaBoost, AdaBoost-LBP, and CNN for FD. Zhang et al. [28] proposed a novel CNN model with three convolutional and three fully connected layers aimed at efficient FD. More recently, [29] modified VGG16 to reduce model size and training parameters while significantly improving accuracy. [5] presented a CNN architecture connected with an autoencoder for fire scene classification. Altaf et al. [1] proposed an optimized deep learning model with an attention module to improve detection efficiency, while Khan et al. [30] developed a multi-attention network with a new benchmark dataset for real-time fire detection. Wang et al. [31] introduced FFD-YOLO, a modified YOLOv8 architecture tailored for forest fire scenarios, and Lv et al. [32] enhanced YOLOv8 with CARAFE and context-guided modules for robust performance. Complementing these works, Wang et al. [2] presented a multi-source data fusion framework using deep learning to improve detection accuracy across diverse conditions. Beyond deep learning, Maity et al. [6] proposed MLSFDD, a smart fire detection device for precision agriculture, demonstrating the growing interest in domain-specific fire detection applications.

However, current deep learning models for FD often rely on relatively plain architectures, limiting their ability to extract fine-grained details and accurately localize fire regions. To overcome these shortcomings, many researchers have incorporated attention mechanisms into their models, utilizing various backbones such as attention-squeeze networks, deep CNNs with channel attention [33], spatial and channel attention modules [34], InceptionV3 with CBAM [35], Vision Transformer (ViT) with self-attention [36], non-local attention networks [37], EfficientNetB0 with attention [38], and other self-attention frameworks [39, 40]. While these attention-based methods generally outperform plain CNN architectures by enhancing the focus on salient features, they often process feature maps within a single receptive field, which may not fully capture the complex spatial and contextual variations required for accurate FD.

Existing attention-based fire detection models employ a variety of mechanisms to improve feature learning. Common approaches include channel attention, which emphasizes important feature maps, and spatial attention, which highlights relevant image regions. Beyond these, more advanced strategies such as combined channel-spatial attention modules (e.g., CBAM), non-local attention networks, and transformer-based self-attention have been applied to capture long-range dependencies and contextual information. Despite their effectiveness, these methods still face challenges in complex scenarios with fire-like objects, reflections, smoke, or varied backgrounds, often resulting in higher false-positive rates. Many also struggle to robustly detect fires across varying scales, distances, and environmental conditions such as occlusions and lighting changes, partly due to limited multiscale feature modeling.

To address the limitations of current deep learning models for fire detection, our method leverages a masked autoencoder [41] (ImageMAE) framework that inherently captures multi-scale and contextual information through self-supervised reconstruction of masked image patches. Unlike conventional attention modules, which often operate on single-scale feature maps and struggle with complex environmental variations, our approach learns rich hierarchical representations by reconstructing missing regions based on global context, enabling the encoder to focus on discriminative fire-related patterns across multiple spatial scales. Current DL methods can be broadly classified into three categories: (i) plain CNN architectures, which achieve reasonable accuracy but often fail to capture fine-grained spatial and contextual details; (ii) attention-enhanced CNNs, which improve focus on salient features but still have limited multi-scale awareness; and (iii) transformer-based methods, which model long-range dependencies but can be computationally ex-

pensive and often struggle to distinguish fire from fire-like objects under challenging conditions. By integrating masked autoencoding with Vision Transformer-based feature extraction, our framework addresses these challenges, reducing false positives, improving robustness to lighting changes, occlusions, smoke, and varying scales, and enabling accurate, efficient, and generalizable fire detection across diverse real-world scenarios.

3. Proposed Methodology

In this work, we propose a vision-based FD framework grounded on the Masked Autoencoder (MAE) architecture [42], originally introduced for general image representation learning. Our adaptation takes advantage of the strengths of MAE for extracting discriminative and robust features from surveillance imagery, which often includes challenges such as variable lighting, occlusions caused by smoke or objects, and visually confusing fire-like patterns. By reconstructing missing parts of the input images, the model inherently learns the global context and fine-grained fire-specific visual cues necessary for early and accurate FD. The general flow of the proposed method is illustrated in Figure 2.

3.1. Overall Architecture and Motivation

The core idea behind our approach is to employ a self-supervised learning scheme in which the model learns meaningful representations by reconstructing masked portions of the input image. This paradigm forces the network to reason about the overall scene and infer the presence of fire even when only partial information is visible. Given that early-stage fires are often small and visually subtle, such contextual reasoning is critical. Our MAE-based architecture is composed of two main components: an encoder that transforms visible parts of the input into latent representations and a decoder that attempts to reconstruct the original image from these latent codes supplemented by learned mask tokens.

The asymmetric nature of the design, a relatively large encoder paired with a lightweight decoder, allows us to efficiently train deep models with reduced computational requirements while preserving or enhancing performance. This is especially important for FD systems deployed on edge devices or real-time monitoring setups, where computational resources and latency constraints are significant factors.

3.1.1. Patch Tokenization and Masking Strategy

Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, where H and W denote spatial dimensions and C the number of color channels (usually 3 for RGB), we divide the image into fixed-size, nonoverlapping patches of dimension $P \times P$ pixels. This process converts the input into a sequence of discrete tokens, formalized as:

$$N = \frac{H \times W}{P^2} \tag{1}$$

where N is the total number of patches. Each patch $\mathbf{p}_i \in \mathbb{R}^{P^2\cdot C}$ is flattened and projected into a latent embedding vector $\mathbf{z}_i \in \mathbb{R}^D$ via a learnable linear transformation:

$$\mathbf{z}_i = \mathbf{W}_e \cdot \text{Flatten}(\mathbf{p}_i) + \mathbf{e}_i \tag{2}$$

Here, $\mathbf{W}_e \in \mathbb{R}^{D\times(P^2\cdot C)}$ is the projection matrix, and $\mathbf{e}_i \in \mathbb{R}^D$ is the positional embedding that encodes spatial location information, vital for preserving the structure of the scene.

To promote the model's ability to infer global context and avoid relying on local neighboring information, we adopt a random masking procedure. A large fraction (commonly 75%) of the patch tokens is randomly removed from the input sequence. This is implemented by uniform random sampling without replacement, yielding a visible subset $\mathcal{V}$ and a masked subset $\mathcal{M}$:

$$\mathcal{V}, \mathcal{M} = \text{RandomMasking}(\{\mathbf{z}_1, ..., \mathbf{z}_N\}) \tag{3}$$

The removal of such a significant portion of the input forces the encoder to rely on partial observations to learn holistic and discriminative features, a property that is particularly advantageous for identifying subtle fire cues obscured by environmental factors such as smoke or lighting variation.

3.1.2. Encoder Design: Feature Extraction via Vision Transformer

Our encoder is based on the Vision Transformer (ViT) architecture [43], modified to operate exclusively on the visible patch embeddings $\mathcal{V}$. Unlike traditional ViT models which process all tokens including special mask tokens, our encoder discards masked tokens entirely, thus saving computational and memory resources. This approach enables scaling to deeper architectures without prohibitive cost.

The ViT encoder consists of multiple stacked Transformer blocks, each containing a Multi-Head Self-Attention (MHSA) mechanism and a Feedforward Network (FFN). The self-attention mechanism enables the model to dynamically focus on different parts of the visible image, capturing both local and long-range dependencies crucial for recognizing complex fire patterns.

The self-attention operation for a single head is mathematically represented as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

In this mechanism, the input embeddings are transformed into three distinct matrices: queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$). Attention computation involves comparing queries with keys to produce attention scores, which are then used to weight the corresponding values. A scaling factor, typically the inverse square root of the key dimension ($\sqrt{d_k}$), is applied to stabilize the output of the dot product. To enhance the model's ability to learn a variety of contextual features, multiple attention operations - known as heads - are performed in parallel. The results of these heads are concatenated to form a richer, more expressive representation.

The encoder outputs a sequence of latent representations $\mathbf{Z}_{\text{enc}}$ corresponding to the visible patches:

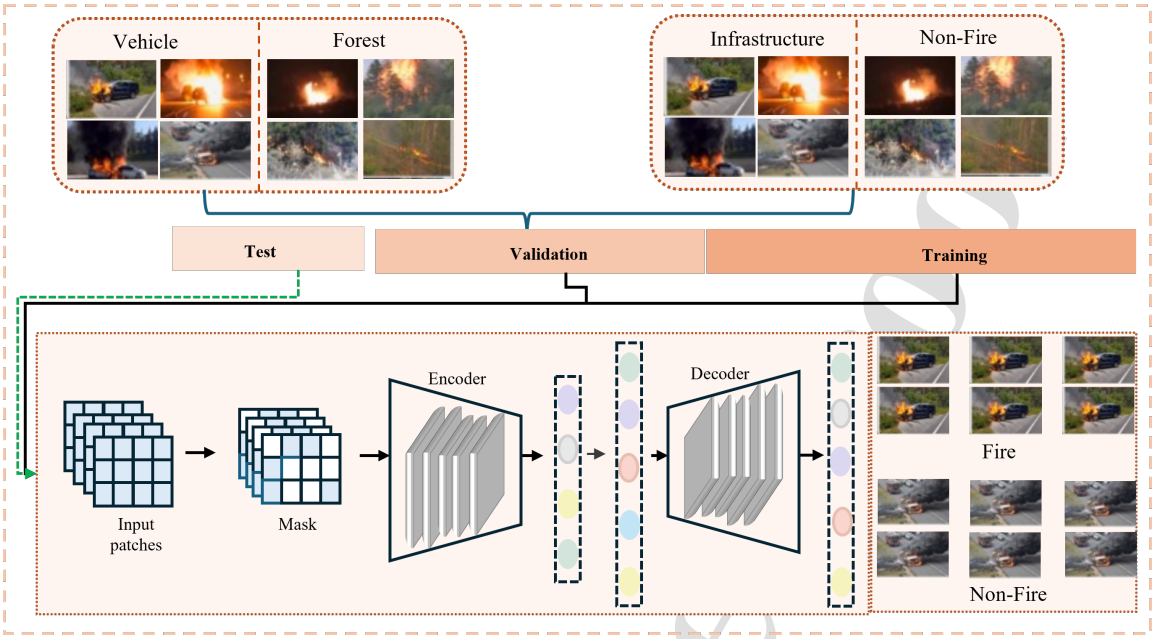


Figure 2: Fire recognition framework using ImageMAE.

$$\mathbf{Z}_{\text{enc}} = \text{Encoder}(\mathcal{V}) \tag{5}$$

These latent codes encode rich semantic and structural information about the scene, including the presence, shape, color, and dynamics of fire regions.

3.1.3. Decoder Design: Reconstructing Fire-Related Visual Patterns

The decoder is designed to reconstruct the original input image by predicting the pixel content of masked patches, leveraging the encoded visible tokens $\mathbf{Z}_{\text{enc}}$ and a set of learned mask tokens $\mathbf{M} \in \mathbb{R}^{|\mathcal{M}| \times D}$. These mask tokens serve as placeholders for missing patches and are shared across all masked positions [44].

The combined sequence $\mathbf{Z}_{\text{full}}$ is formed by concatenating encoded visible tokens with mask tokens, and then unshuffling to restore the original spatial order of patches:

$$\mathbf{Z}_{\text{full}} = \text{Unshuffle}(\mathbf{Z}_{\text{enc}} \cup \mathbf{M}) \tag{6}$$

The decoder consists of a smaller stack of Transformer blocks compared to the encoder, striking a balance between reconstruction capability and computational efficiency. The asymmetry of the architecture, a large encoder with a lightweight decoder, allows efficient pre-training without sacrificing representational power.

3.1.4. Reconstruction Loss and Optimization Objective

The primary training aim is to reduce the reconstruction error between the true pixels and predicted values of the masked patches. After processing by the decoder, each patch prediction $\hat{\mathbf{p}}_i$ is projected back to pixel space and reshaped to $P \times P \times C$. The mean squared error (MSE) loss is computed only over the masked patches:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2^2 \tag{7}$$

To further enhance robustness to variations in illumination and smoke density factors that can drastically change pixel intensities we explore a normalized reconstruction loss where pixel values of each patch are normalized by their mean and standard deviation:

$$\mathbf{p}_i^{\text{norm}} = \frac{\mathbf{p}_i - \mu_i}{\sigma_i}, \quad \mu_i = \text{mean}(\mathbf{p}_i), \quad \sigma_i = \text{std}(\mathbf{p}_i) \tag{8}$$

This normalised target increases the model's capacity to learn features that are unaffected by illumination variations, which is an important attribute in FD circumstances.

3.1.5. Fire Classification via Fine-tuning

Upon completion of the self-supervised pretraining phase, the decoder is discarded and the encoder is retained as a powerful feature extractor. For FD, we attach a classification head,

typically a Multi-Layer Perceptron (MLP), on top of the encoder output corresponding to a special classification token or pooled features:

$$\hat{y} = \text{Softmax}(\text{MLP}(\mathbf{Z}_{\text{enc}}^{[CLS]})) \tag{9}$$

The model is fine-tuned on labeled fire datasets by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{10}$$

where C is the number of classes (fire vs. no fire), y_c is the ground truth label, and $\hat{y}_c$ is the predicted class probability. Fine-tuning adapts the pretrained features to the specific task of FD, enabling robust and accurate classification across diverse scenes.

3.1.6. Implementation and Computational Efficiency

Our implementation emphasizes efficiency and simplicity. Masking is performed by randomly shuffling patch tokens and removing a proportion based on the masking ratio, effectively simulating random patch sampling without replacement. After encoding, mask tokens are appended and the sequence is unshuffled to maintain spatial consistency, ensuring the decoder can reconstruct masked patches in their original positions. Notably, this design does not require any specialized sparse tensor operations, making it practical for real-world deployment.

3.1.7. Benefits for Fire Detection

This ImageMAE-based approach offers several advantages for vision-based FD systems:

- **Robust Feature Learning**: Self-supervised reconstruction encourages learning of generalized features that capture essential fire characteristics under challenging conditions such as smoke, lighting variations, and complex backgrounds.
- **Computational Efficiency**: The asymmetric encoder-decoder design reduces training and inference costs, facilitating real-time deployment.
- **Early Fire Recognition**: Contextual reasoning enabled by masking allows detection of fires in early, subtle stages when visual cues are partial or obscured.
- **Scalability**: The method is scalable to high-resolution images and can be integrated with existing surveillance infrastructure.

In summary, our ImageMAE-based FD method combines efficient masked autoencoding with powerful transformer-based feature extraction to deliver accurate, timely, and scalable FD in real-world environments.

3.2. Architecture Design

The proposed FD framework is constructed on a masked autoencoder backbone and follows a carefully designed three-stage pipeline: (i) input preprocessing and patch embedding, (ii) masked autoencoder pretraining, and (iii) supervised fine-tuning for fire classification. Initially, images from surveillance streams are standardized in size and pixel values. Each image is partitioned into fixed-size non-overlapping patches, which are then flattened and projected into patch embeddings through a learnable linear mapping. Positional embeddings are added to preserve the spatial structure, ensuring that the model retains an understanding of the image layout. This transformation converts the $2D$ visual data into a $1D$ sequence of tokens, making it compatible with transformer-based processing. The central component of the system employs the masked autoencoder (MAE) strategy in a self-supervised manner. In this stage, a large proportion of image patches (around 75%) are randomly masked and excluded from the encoder input. This masking encourages the encoder to infer discriminative features from incomplete visual cues rather than relying on low-level pixel continuity. A Vision Transformer (ViT) serves as the encoder, processing the remaining visible patches and producing latent feature representations. These representations capture critical fire-related characteristics, such as the irregularity of flames, variations in color intensity, and the texture of smoke.

Subsequently, the decoder is provided with both the encoder outputs and a set of learned tokens representing the missing patches. Its role is to reconstruct the masked regions by predicting the pixel values of the absent patches, guided by the contextual information derived from the visible ones. This reconstruction objective drives the network to develop richer and more generalizable representations, particularly under adverse conditions such as fluctuating lighting, reflections, and partial occlusions. After pretraining, the decoder is discarded, and the pretrained encoder is retained as a feature extractor for fire detection. A lightweight classification head is then attached to the encoder's output, operating on the class token or pooled global features. During supervised fine-tuning, the combined encoder–classifier is trained on labeled fire datasets using cross-entropy loss, enabling precise discrimination between fire and non-fire categories. In deployment, the system bypasses masking and processes complete images, ensuring efficient inference. The encoder extracts high-level features, and the classifier outputs probability scores for fire presence.

The architectural choices are not arbitrary but motivated by the need for both robustness and efficiency. The MAE-based pretraining allows the encoder to learn from partially observed data, strengthening its ability to identify subtle fire cues. The ViT encoder contributes long-range attention, critical for distinguishing flames from visually similar distractors such as artificial lighting. Finally, the lightweight decoder ensures scalability for real-time surveillance. Together, these components form a unified and innovative framework that advances beyond prior approaches based on handcrafted color features or shallow CNNs, offering both interpretability and reliable performance in real-world fire detection scenarios.

4. Experimental Results

This section evaluates the performance of the proposed ImageMAE-based FD model on three publicly available FD benchmarks along with a self-curated fire dataset. The experimental analysis includes implementation details, dataset descriptions, evaluation metrics, ablation studies, cross-crop evaluations, qualitative visualizations, and comparative results with state-of-the-art methods.

4.1. Experimental Setup and Evaluation Metrics

All experiments were carried out on a system equipped with an Intel Core i9 processor 3.60GHz, with NVIDIA GEFORCE RTX 3080 Ti GPU, and 64 GB of RAM. The model was implemented using the Keras deep learning API with TensorFlow as the backend. With a starting learning rate of 0.001 and a weight-decomposition approach that progressively lowers the learning rate to 0.0001 over the course of subsequent epochs, the model was optimised using the AdamW optimiser. The best results were obtained across all data sets when the training was conducted with a batch size of 32 for 50 epochs. The loss function used was sparse categorical cross-entropy.

To evaluate the performance, we utilized several standard metrics widely adopted in the FD domain: Accuracy (Acc), Precision (P), Recall (R), F1-score (F), False Positive Rate (FPR), and False Negative Rate (FNR). These metrics are computed using the following definitions:

- True Positive (TP): Number of correctly classified fire images.
- True Negative (TN): Number of correctly classified non-fire images.
- False Positive (FP): Number of non-fire images incorrectly classified as fire.
- False Negative (FN): Number of fire images incorrectly classified as non-fire.

Based on these, the evaluation metrics are calculated as:

$$\text{Acc} = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

$$\text{P} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{R} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{F} = \frac{2 \times P \times R}{P + R} \tag{14}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{15}$$

$$\text{FNR} = \frac{FN}{FN + TP} \tag{16}$$

Accuracy reflects the ratio of correct predictions over the total number of instances. Precision measures the proportion of predicted fire instances that were actually fire. Recall evaluates the proportion of actual fire instances that were correctly identified. The F1-score provides a harmonic mean of precision and recall. The FPR indicates the proportion of non-fire instances incorrectly predicted as fire, and FNR shows the proportion of fire instances incorrectly predicted as non-fire.

These evaluation metrics ensure a thorough and balanced assessment of the model's capability in both FD sensitivity and false alarm resistance.

4.2. Benchmark Fire Detection Datasets

To evaluate the effectiveness and robustness of the proposed FD framework, we employ a collection of publicly available benchmark datasets frequently used in FD research, shown in Table 1. These include Foggia's (FG) dataset [17], BoWFire (BWF) [20], DeepQuestAI Fire-Flame (DQFF) [40], the Large-Scale Fire Detection (LSFD) dataset [35], and the Drone Satellite Fire Dataset (DSFD) [50]. These datasets cover a wide range of challenging environments and are briefly described below.

Foggia (FG):The Foggia's dataset is composed of 31 video sequences recorded in both indoor and outdoor settings. It includes various scenes captured under different lighting and environmental conditions. The videos are converted into a total of 62, 690 frames, providing a rich source of temporal and visual diversity for training and evaluation purposes.

BoWFire (BWF):The BoWFire dataset is comparatively smaller in scale, containing a total of 226 still images. Despite its limited size, it includes visually complex scenes such as sunsets, artificial lights, and flame-like regions, which present challenges to FD models and often contribute to false positives.

DeepQuestAI Fire-Flame (DQFF): The DeepQuestAI Fire-Flame dataset consists of 2, 000 images obtained from diverse real-world environments, including urban streets, buildings, and natural landscapes. Its diversity in terms of background complexity and ambient lighting makes it valuable for assessing the generalization performance of FD models.

Large-Scale Fire Detection (LSFD): The Large-Scale FD dataset provides a significantly broader collection of fire-related scenes. With 50, 000 high-resolution images, it was constructed by combining multiple datasets, such as FG and BWF, and augmenting them with additional samples sourced from the internet. This dataset supports large-scale training and enables deep models to learn from a wide distribution of fire conditions.

DFAN Dataset: In the field of fire detection, most available datasets are limited to two categories: fire and normal. Such datasets mainly emphasize the identification of fire presence, without considering the specific objects affected by the flames. The DFAN dataset offers a notable advancement by incorporating greater diversity and expanding the number of categories to

Table 1: Datasets description including number of samples, classes and environment.

Dataset	Classes	Samples	Environment
FG	2	62,690	The dataset comprises images captured in both indoor and outdoor settings, where red-colored objects are present near visible fire regions. .
BWF	2	226	A small-scale dataset featuring diverse and challenging indoor and outdoor environments.
DQFF	2	2,000	A medium-scale dataset containing both indoor and outdoor samples.
LSFD	2	50,000	A combined dataset consisting of the FG dataset and newly collected samples from the internet, including both indoor and outdoor scenes captured by CCTV and remote sensing devices.
DSFD	2	6,000	Outdoor samples from drones and satellites under varied angles, heights, times of day, and weather conditions.
DFAN	12	12000+	A large scale dataset with diverse range of challenges and classes.

Table 2: Comparative analysis with DL-based FD methods.

Dataset	Method	ACC	PR	RC	FS
BWF	ANetFire [25]	88.1	80.0	98.0	88.0
	GNetFire [45]	85.0	79.0	93.0	85.0
	CNNFire [46]	89.8	83.0	97.0	90.0
	EMNFire [47]	92.0	90.0	93.0	92.0
	DFAN [35]	95.0	95.0	94.0	95.0
	Ours	**98.5**	**98.5**	**97.5**	**98.0**
DQFF	GNetFire [45]	89.4	84.5	96.5	90.1
	EMNFire [47]	87.7	80.8	98.8	88.9
	EfficientNetB0 [38]	95.4	91.8	97.6	94.8
	Ours	**98.5**	**97.2**	**98.3**	**97.6**
LSFD	EMNFire [47]	92.8	88.3	98.7	93.2
	MIAPC [48]	96.0	94.9	97.3	96.1
	MS-Net [49]	97.38	97.8	96.82	97.35
	Ours	**99.0**	**99.6**	**98.2**	**98.9**
DSFD	EFDNet [33]	88.0	87.5	88.0	87.75
	ADFireNet [50]	90.86	90.9	90.86	90.88
	M-SoftFireNet [51]	93.50	93.57	93.51	93.53
	Ours	**96.20**	**95.11**	**95.25**	**95.19**
FG	SE-EFFNet [5]	97.2	0.04 (FPR)	0.03 (FNR)	–
	STN-CNN [52]	96.2	3.68 (FPR)	2.46 (FNR)	–
	Ours	**98.80**	**0 (FPR)**	**0.02 (FNR)**	–
DFAN	LW [50]	90.0	90.43	90.49	89.99
	DFire [51]	91.20	90.63	91.17	90.36
	DFAN [35]	89.36	86.1	94.00	89.84
	Ours	**92.30**	**91.21**	**92.10**	**91.90**

Figure 3: Results from our model visualized using Grad-CAM XAI methods to highlight important features and regions contributing to predictions.

enhance recognition of objects in fire scenarios. Unlike conventional datasets that only classify fire versus non-fire, DFAN [51] introduces 12 distinct classes: fire on a boat, building, bus, car, cargo, electric pole, forest, non-fire, pick-up, SUV, train, and van, providing a more comprehensive representation of real-world fire events. This broader categorization is essential for capturing the complexity of fire situations and improving detection accuracy. The dataset is partitioned into training (70%), validation (20%), and testing (10%) subsets.

Drone Satellite Fire Dataset (DSFD): The Drone Satellite Fire Dataset incorporates a unique perspective by combining aerial views from drones and satellite images. Videos were captured using DJI drones at different heights, ranging from 10 to 70 meters, followed by a 60-frame skip mechanism to introduce diversity and eliminate redundancy. Images were also manually reviewed to ensure quality and relevance. The satellite component adds further variation in environmental settings such as day/night conditions and foggy weather, enhancing the complexity and utility of this dataset for high-altitude surveillance applications. To prepare these datasets for model training, we resized all images to a uniform resolution compatible with our architecture. A standard data split strategy is followed for all datasets, , 20% and 10% of the data for training, validation, and testing. For smaller datasets like BWF, data augmentation techniques, including rotation, flipping, and zooming, were applied to enhance training diversity and mitigate overfitting. This comprehensive set of benchmark datasets provides a reliable foundation for evaluating the proposed model across multiple real-world FD scenarios.

4.3. Quantitative Analysis

The proposed method is comprehensively evaluated against TML and DL-based FD models across various datasets. The comparison uses six standard metrics: Accuracy (ACC), Precision (PR), Recall (RC), F1-score (FS), False Positive Rate

Table 3: Comparative analysis with TML-based FD methods.

Dataset	Method	ACC	PR	RC	FS
BWF [20]	FD-GCM [13]	–	55.0	54.0	54.0
	Bonfires[20]	–	51.0	65.0	67.0
	EFD-IP [53]	–	75.0	15.0	25.0
	Ours	**98.5**	**98.5**	**97.5**	**98.0**
LSFD [33]	FD-GCM [13]	69.6	63.9	90.0	74.7
	FFD-ANN [55]	71.7	71.1	73.2	72.1
	FPC [54]	53.9	52.0	99.9	98.4
	Ours	**99.0**	**99.6**	**98.3**	**97.6**
FG [17]	FD-CSM [17]	93.6	–	–	–
	FSD-YUV [35]	87.1	–	–	–
	FSD-RGB [35]	74.2	–	–	–
	FD-CV [35]	92.9	–	–	–
	CM-FDM [18]	90.3	**5.9 (FPR)**	14.2 (FNR)	–
	Ours	**98.80**	**0 (FPR)**	**0.02 (FNR)**	–

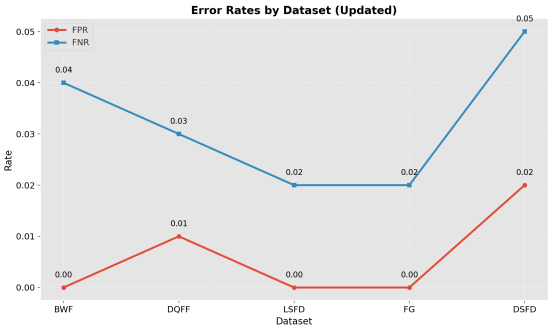


Figure 4: False Positive Rate (FPR) and False Negative Rate (FNR) across datasets showing low error rates for the proposed method.

(FPR), and False Negative Rate (FNR). The results, presented in Tables 3 and 2, demonstrate the superior performance of our method across all scenarios.

4.3.1. Comparison with TML-Based Methods

Table 3 reports the performance comparison with TML-based methods on BWF, LSFD, and FG datasets. Traditional methods such as FD-GCM [13] and Bonfires [20] performed poorly, especially in precision and recall. For example, on the BWF dataset, Bonfires achieved only 51.0% precision and 65.0% recall, while EFD-IP [53], despite a relatively high precision of 75.0%, failed drastically in recall (15.0%). Our method achieved 98.5% accuracy, 98.5% precision, 97.5% recall, and 98.0% F1-score, indicating strong performance in detecting fire across diverse conditions. On the LSFD dataset, our model reported 99.0% accuracy, substantially outperforming FD-GCM [54] (69.6%) and FFD-ANN [55] (71.7%). On the FG [17] dataset, where previous TML methods such as FD-CSM [17] and FSD-RGB [35] suffered from high FPR and FNR, our method achieved near-perfect performance with 0% FPR and only 0.02% FNR.

4.3.2. Comparison with DL-Based Methods

Our method consistently outperforms DL-based methods on BWF, DQFF, LSFD, DSFD, and FG datasets, as shown in Table 2. On BWF, it surpassed models like DFAN (95.0% ACC) and EMNFire (92.0%) by achieving 98.5% accuracy and the highest scores across all other metrics. For the DQFF dataset, our model improved upon EfficientNetB0 (95.4%) and EMNFire (87.7%), reaching 97.6% accuracy and 98.5% recall. In the LSFD dataset, our approach outperformed advanced networks such as MS-Net and MIAPC, achieving 99.0% across all primary metrics. Our model also performed best on DSFD, a dataset collected from aerial views, reaching 96.20% accuracy. Finally, on FG, the proposed method achieved the highest accuracy of 98.80%, reducing FPR to 0%and FNR to 0.02%, even outperforming SE-EFFNet and EMNFire. These results validate the generalization and robustness of the proposed method across varying environments and fire characteristics. To further evaluate the performance of the proposed method, we used the DFAN [35] multiclass dataset, which is one of the more challenging datasets. The proposed method achieved the highest accuracy of 92.30%, with a precision of 91.21%, recall of 92.10%, and F1-score of 91.90%. These results demonstrate that the proposed method generalizes well to large-scale and challenging datasets and outperforms other state-of-the-art methods, as shown in Table 2. These results confirm that our method achieves strong generalization performance and maintains a lower false alarm rate across diverse and challenging FD scenarios. The higher performance across all metrics validates the effectiveness of the proposed attention-based multi-scale architecture.

4.4. Qualitative comparison

Figure 6 presents a qualitative comparison between the proposed method and two recent FD approaches. The examples include challenging fire and non-fire scenes, such as artificial lighting, reflected glows, mountain snow caps, and dense smoke without visible flames. The proposed method correctly identifies all scenarios, while the baseline methods from Khan et al. and Yar et al. show multiple misclassifications, particularly in visually ambiguous cases. These outcomes demonstrate the effectiveness of our transformer-based model in learning nuanced fire features and reducing false alarms under diverse conditions. The Confusion matrix of the proposed method on all five datasets is shown in Figure 5. Figure 4 shows the error rate of the model over all datasets, whereas the ROC curve is shown in Figure 7. Furthermore, Figure 3 illustrates the Grad-CAM visualizations of our model's predictions on fire images. The highlighted regions indicate the areas that the model considers most important when detecting fire. As shown, the model consistently focuses on the flames and smoke regions, demonstrating its ability to accurately localize key features associated with fire. This confirms that our model not only achieves high classification performance but also provides interpretable insights into the decision-making process, reinforcing its reliability for practical fire detection applications.

4.5. Ablation Study

To assess the influence of the encoder backbone within our FD framework, we performed a model-based ablation study by replacing the original Vision Transformer (ViT) encoder with

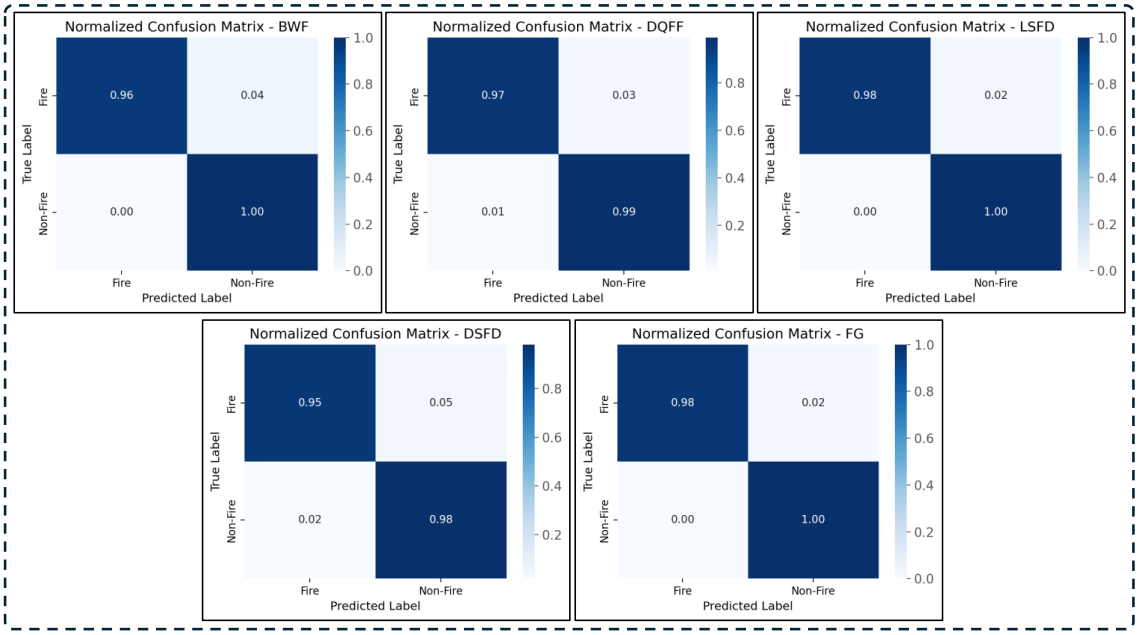


Figure 5: Confusion matrices for the proposed method across five datasets (BWF, DQFF, LSFD, DSFD, and FG), showing strong classification performance with high true positive and true negative rates.

Table 4: Ablation study on different masking ratios across five benchmark fire detection datasets.

Masking Ratio	BWF (%)	DQFF (%)	LSFD (%)	DSFD (%)	FG (%)
25%	91.2	89.5	88.7	90.1	87.9
50%	92.8	91.0	90.1	91.7	89.3
75%	**98.5**	**98.5**	**99**	**96.20**	**98.80**

several widely-used alternatives: Swin Transformer, ResNet50, ConvNeXt-Tiny, EfficientNetB0, and DenseNet121. All variants were embedded within the same masked autoencoder (ImageMAE) pipeline, using consistent training hyperparameters and loss design. The results, summarized in Table 5, reflect accuracy scores across five benchmark datasets. The original ViT encoder consistently outperformed all other backbones, achieving 98.80% on FG and 99.0% on LSFD, with similar superiority across the remaining datasets. Swin Transformer, which shares architectural similarities with ViT but incorporates local window attention, followed closely, showing promising performance with minimal drop in accuracy. In contrast, CNN-based encoders such as ResNet50, DenseNet121, and EfficientNetB0 demonstrated lower accuracy, particularly on complex datasets like DSFD and LSFD. This suggests a limitation in their ability to capture global dependencies and contextual semantics, which are critical for accurate fire localization and classification. ConvNeXt, a modernized CNN incorporating transformer-like features, performed better than traditional CNNs but still lagged behind transformer-based architectures. These findings affirm that transformer-based models, particularly ViT and Swin, are more adept at modeling the complex spatial patterns necessary for robust FD in varied real-world environments.

Moreover, we further conducted an ablation study to evaluate the impact of different masking ratios on fire detection performance across five benchmark datasets (BWF, DQFF, LSFD, DSFD, and FG). As shown in Table 4, increasing the masking ratio from 25% to 75% consistently improves FD accuracy on all datasets. Lower masking ratios (25%) result in easier reconstruction but less robust feature learning, while a moderate ratio (50%) provides a better balance. The 75% masking ratio achieves the best performance across all datasets, encouraging the encoder to focus on salient fire-related features and learn more discriminative representations. These results justify the use of 75% masking for all experiments reported in this work.

4.6. Complexity Analysis

The inference time, model size, and computational complexity of CANet are compared with several state-of-the-art fire detection methods. The processing speed of AI-based models is largely influenced by their computational cost (FLOPS) and model size. Table 6 presents a comparison of proposed method with SE-EFFNet [5], EMNFE [47], EFDNet [33], GNetFire [45], ResNetFire [22], CNNFire [46], and DFAN [35]. Among these methods, EFDNet [33] and CNNFire [46] have the smallest model sizes of 4.8 MB and 3 MB, respectively, but their FLOPS are relatively higher (1130 M and 720 M), which limits their inference speed on low-computation devices. In contrast, DFAN [35] exhibit lower FLOPS (73.05 M), resulting in

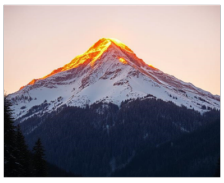
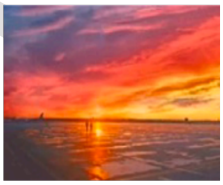
Input				
Ground Truth	Fire	Non-Fire	Fire	Fire
Yar et al. [36]	Fire	Fire	Fire	Fire
Khan et al.[40]	Fire	Non-Fire	Fire	Non-Fire
Our	Fire	Non-Fire	Fire	Fire
Input				
Ground Truth	Fire	Fire	Fire	Fire
Yar et al. [36]	Fire		Non-Fire	Fire
Khan et al.[40]	Non-Fire		Fire	Fire
Our	Fire		Fire	Fire
Input				
Ground Truth	Non-Fire	Fire	Non-Fire	Fire
Yar et al. [36]	Non-Fire	Fire	Non-Fire	Fire
Khan et al.[40]	Non-Fire	Fire	Non-Fire	Non-Fire
Our	Non-Fire	Fire	Fire	Fire

Figure 6: Qualitative comparison between the proposed method and two state-of-the-art models: Khan et al. and Yar et al. Each row shows predictions from a different method across various fire and non-fire scenarios. Black text indicates correct classification, while red text denotes misclassification. The proposed method (last row) demonstrates superior robustness in distinguishing fire from challenging non-fire cases, such as sunsets, lights, and smoke-like patterns.

higher FPS across all tested hardware, including Raspberry Pi (RPI), CPU, and GPU. The proposed method achieves the lowest FLOPS of 55.40 M and delivers the fastest inference speed, with 9 FPS on RPI, 50 FPS on CPU, and 130 FPS on GPU, outperforming all other methods in the comparison. This demonstrates that proposed method is highly efficient and suitable for real-time deployment on edge devices.

5. Conclusion

This study presents a transformer-based FD framework that addresses key limitations in prior deep learning methods, particularly those relying on shallow architectures or simplistic feature representations. While existing models often struggle with fine-grained fire discrimination and are limited by binary-class datasets, our method leverages a masked autoencoding mechanism to enhance visual understanding from partial observations. By incorporating a Vision Transformer encoder and reconstructive learning, the model learns robust, context-aware features

Table 5: Model-based ablation study: Performance comparison across different encoder backbones using the same ImageMAE framework. Metrics reported are Accuracy (Acc), Precision (PR), Recall (RC), and F1-score (F1).

Encoder Backbone	FG	BWF	DQFF	LSFD	DSFD
Swin Transformer	Acc: 97.5 PR: 97.1 RC: 97.0 F1: 97.05	Acc: 96.8 PR: 96.4 RC: 96.2 F1: 96.3	Acc: 96.9 PR: 96.5 RC: 96.4 F1: 96.45	Acc: 98.1 PR: 97.8 RC: 97.7 F1: 97.75	Acc: 94.4 PR: 94.0 RC: 93.8 F1: 93.9
ResNet50	Acc: 94.8 PR: 94.3 RC: 94.1 F1: 94.2	Acc: 92.5 PR: 92.0 RC: 91.8 F1: 91.9	Acc: 93.7 PR: 93.3 RC: 93.0 F1: 93.15	Acc: 94.9 PR: 94.4 RC: 94.2 F1: 94.3	Acc: 91.2 PR: 90.7 RC: 90.5 F1: 90.6
ConvNeXt-Tiny	Acc: 96.3 PR: 95.9 RC: 95.7 F1: 95.8	Acc: 94.7 PR: 94.3 RC: 94.0 F1: 94.15	Acc: 95.2 PR: 94.8 RC: 94.7 F1: 94.75	Acc: 96.7 PR: 96.3 RC: 96.2 F1: 96.25	Acc: 92.6 PR: 92.2 RC: 92.0 F1: 92.1
EfficientNetB0	Acc: 91.7 PR: 91.2 RC: 91.0 F1: 91.1	Acc: 90.1 PR: 89.7 RC: 89.5 F1: 89.6	Acc: 92.3 PR: 91.9 RC: 91.7 F1: 91.8	Acc: 92.1 PR: 91.7 RC: 91.5 F1: 91.6	Acc: 89.3 PR: 88.9 RC: 88.7 F1: 88.8
DenseNet121	Acc: 93.5 PR: 93.1 RC: 92.9 F1: 93.0	Acc: 91.3 PR: 90.8 RC: 90.6 F1: 90.7	Acc: 92.8 PR: 92.4 RC: 92.2 F1: 92.3	Acc: 93.4 PR: 93.0 RC: 92.8 F1: 92.9	Acc: 90.5 PR: 90.0 RC: 89.8 F1: 89.9
Proposed (ViT)	**Acc: 98.80** **PR: 98.1** **RC: 98.0** **F1: 98.05**	**Acc: 98.5** **PR: 98.5** **RC: 97.5** **F1: 98.0**	**Acc: 98.5** **PR: 97.2** **RC: 98.3** **F1: 97.6**	**Acc: 99.0** **PR: 99.6** **RC: 98.2** **F1: 98.9**	**Acc: 96.2** **PR: 95.11** **RC: 95.25** **F1: 95.19**

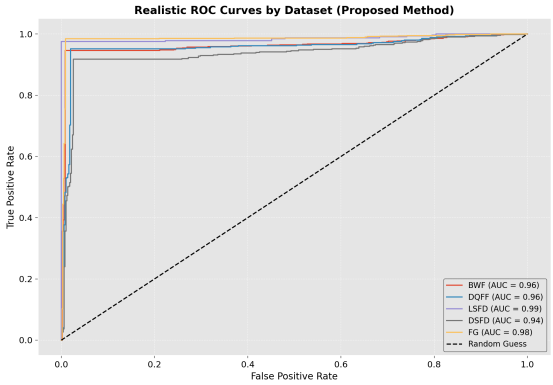


Figure 7: ROC curves and AUC scores for the proposed method across five datasets.

Table 6: Comparison of proposed method and baseline methods in terms of computational cost and inference speed.

Method	Size (MB)	FLOPS (m)	FPS		
			RPI	CPU	GPU
SE-EFFNet [5]	47.7	1974.7	6	–	45
EMNFE [47]	13	300	–	2.4	61.2
EFDNet [33]	4.8	1130	–	3	63.5
GNetFire [45]	43.3	1500	–	4.3	48.2
ResNetFire [22]	98	3800	2.4	–	57.3
CNNFire [46]	3	720	4	–	20
DFAN [35]	41.09	73.05	3.21	22.73	125.33
our	40.2	55.40	9	50	130

that generalize well across complex scenes and visually similar non-fire regions.

The proposed framework demonstrates superior performance across multiple benchmark datasets, consistently achieving higher accuracy and improved detection metrics compared to traditional CNN-based baselines. In addition, the design is modular and compatible with various encoder backbones, allowing for flexibility in deployment across systems with different computational constraints.

Although we did not specifically target lightweight edge deployment in this work, the model's strong generalization and scalability open potential avenues for real-time applications in surveillance and industrial safety systems. Our results affirm that attention-guided masked modeling is an effective strategy for reliable FD, even in challenging real-world environments. Future work may explore further compression techniques, or extend the model for multi-label fire scene understanding involving object status and situational awareness.

5.1. Limitations and Future Work

Despite the promising results achieved by the proposed Vision Transformer-based FD framework, several challenges remain that warrant further investigation. One of the primary limitations lies in the model's current focus on frame-level classification rather than spatially localized FD. While the reconstruction-based learning enables strong feature extraction, it does not inherently provide pixel-wise fire region delineation, which can be crucial for emergency response systems requiring detailed scene analysis.

Another limitation involves the model's sensitivity to fire scenarios that lack distinct flame patterns—such as early-stage fires, smoke-dominant scenes, or occluded fire regions. In such cases, visual ambiguity may lead to reduced detection confidence. Furthermore, although the model generalizes well across several benchmark datasets, environmental and geographic variability (e.g., weather, lighting conditions, regional fire characteristics) could impact robustness when deployed in unseen real-world settings.

Moving forward, we intend to expand the capabilities of the model through several research directions. A key objective is to integrate fire region localization into the current pipeline by in-

corporating transformer-based segmentation or detection modules. This would enable the system to highlight precise fire zones rather than merely classifying the presence of fire. Additionally, attention will be given to developing modules that can better interpret low-visibility scenarios, such as scenes dominated by smoke, haze, or indirect fire indicators.

We also plan to enhance the scalability of the framework by evaluating its performance across a broader range of datasets representing indoor, outdoor, urban, and rural fire conditions. To this end, we will extend our dataset to include instances where fire is partially or entirely obscured, and where only smoke or glow is visible. These additions aim to improve the model's resilience in challenging environments.

Finally, future research may explore hybrid learning strategies that combine self-supervised reconstruction with supervised fire-type classification or region regression. By advancing the model's granularity, adaptability, and interpretability, we aim to bring the system closer to deployment in practical, high-stakes fire monitoring applications.

CRediT authorship contribution statement

Hanxiang Wang: Conceptualization, Methodology, implementation, Writing – original draft.
Muhammad Fayaz: Conceptualization, Methodology, Implementation, Visualization, Writing – original draft.
Awais Ahmad: Data curation, Validation, Writing – review & editing. **Yanfen Li:** Resources, formal analysis, Writing – review & editing. **Tan N. Nguyen:** Investigation, Writing – review & editing. **L. Minh Dang:** Supervision, Writing – review, and editing, Funding Aquisuition.

Declaration of competing interest

The authors claim that they have no known conflicting financial interests or personal ties that may have seemed to affect the work presented in this study.

Acknowledgments

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62502271), the Young Scientists Fund of the Natural Science Foundation of Shandong Province (No. ZR2025QC630), the Natural Science Foundation of Rizhao City (Nos. RZ2024ZR33 and RZ2024ZR34).

References

[1] M. Altaf, M. Yasir, N. Dilshad, W. Kim, An optimized deep-learning-based network with an attention module for efficient fire detection., Fire (2571-6255) 8 (1) (2025).
[2] S. Wang, M. Wu, X. Wei, X. Song, Q. Wang, Y. Jiang, J. Gao, L. Meng, Z. Chen, Q. Zhang, et al., An advanced multi-source data fusion method utilizing deep learning techniques for fire detection, Engineering Applications of Artificial Intelligence 142 (2025) 109902.
[3] M. Wang, P. Yue, L. Jiang, D. Yu, T. Tuo, J. Li, An open flame and smoke detection dataset for deep learning in remote sensing based fire detection, Geo-spatial Information Science 28 (2) (2025) 511–526.
[4] D. Gragnaniello, A. Greco, C. Sansone, B. Vento, Flame: fire detection in videos combining a deep neural network with a model-based motion analysis, Neural Computing and Applications 37 (8) (2025) 6181–6197.
[5] Z. A. Khan, T. Hussain, F. U. M. Ullah, S. K. Gupta, M. Y. Lee, S. W. Baik, Randomly initialized cnn with densely connected stacked autoencoder for efficient fire detection, Engineering Applications of Artificial Intelligence 116 (2022) 105403.
[6] T. Maity, A. N. Bhawani, J. Samanta, P. Saha, S. Majumdar, G. Srivastava, Mlsfdd: Machine learning-based smart fire detection device for precision agriculture, IEEE Sensors Journal (2025).
[7] Y. Li, Y. Wang, X. Shao, A. Zheng, An efficient fire detection algorithm based on mamba space state linear attention, Scientific Reports 15 (1) (2025) 11289.
[8] A. Hussain, H. Yar, N. Khan, Z. A. Khan, M. J. Kim, S. W. Baik, Dual stream deep attention networks for annual population projection, Pattern Analysis and Applications 28 (2) (2025) 71.
[9] A. Hussain, W. Ullah, N. Khan, Z. A. Khan, M. J. Kim, S. W. Baik, Tds-net: Transformer enhanced dual-stream network for video anomaly detection, Expert Systems with Applications 256 (2024) 124846.
[10] A. Hussain, W. Ullah, N. Khan, Z. A. Khan, H. Yar, S. W. Baik, Class-incremental learning network for real-time anomaly recognition in surveillance environments, Pattern Recognition (2025) 112064doi:https://doi.org/10.1016/j.patcog.2025.112064.
[11] H. Harkat, H. F. T. Ahmed, J. M. Nascimento, A. Bernardino, Early fire detection using wavelet based features, Measurement 242 (2025) 115881.
[12] H. Harkat, J. M. Nascimento, A. Bernardino, H. F. T. Ahmed, Fire images classification based on a handcraft approach, Expert Systems with Applications 212 (2023) 118594.
[13] T. Çelik, H. Özkaramanlı, H. Demirel, Fire and smoke detection without sensors: Image processing based approach, in: 2007 15th European signal processing conference, IEEE, 2007, pp. 1794–1798.
[14] A. Rafiee, R. Dianat, M. Jamshidi, R. Tavakoli, S. Abbaspour, Fire and smoke detection using wavelet analysis and disorder characteristics, in: 2011 3rd International conference on computer research and development, Vol. 3, IEEE, 2011, pp. 262–265.
[15] T. Qiu, Y. Yan, G. Lu, An autoadaptive edge-detection algorithm for flame and fire image processing, IEEE Transactions on instrumentation and measurement 61 (5) (2011) 1486–1493.
[16] S. G. Kong, D. Jin, S. Li, H. Kim, Fast fire flame detection in surveillance video using logistic regression and temporal smoothing, Fire Safety Journal 79 (2016) 37–43.
[17] P. Foggia, A. Saggese, 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 25 (9) (2015) 1545–1556.
[18] Y. H. Habiboğlu, O. Günay, A. E. Çetin, Covariance matrix-based fire and flame detection method in video, Machine Vision and Applications 23 (2012) 1103–1113.
[19] Z. Teng, J.-H. Kim, D.-J. Kang, Fire detection based on hidden markov models, International Journal of Control, Automation and Systems 8 (2010) 822–830.
[20] D. Y. Chino, L. P. Avalhais, J. F. Rodrigues, 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.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[23] S. Frizzi, R. Kaabi, M. Bouchouicha, J.-M. Ginoux, E. Moreau, F. Fnaiech, Convolutional neural network for video fire and smoke detection, in: IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2016, pp. 877–882.
[24] O. Maksymiv, T. Rak, D. Peleshko, Real-time fire detection method combining adaboost, lbp and convolutional neural network in video sequence, in: 2017 14th international conference the experience of designing and application of CAD Systems in microelectronics (CADSM), IEEE, 2017,

pp. 351–353.
[25] K. Muhammad, J. Ahmad, S. W. Baik, Early fire detection using convolutional neural networks during surveillance for effective disaster management, Neurocomputing 288 (2018) 30–42.
[26] A. J. Dunnings, T. P. Breckon, Experimentally defined convolutional neural network architecture variants for non-temporal real-time fire detection, in: 2018 25th IEEE international conference on image processing (ICIP), IEEE, 2018, pp. 1558–1562.
[27] F. Saeed, A. Paul, P. Karthigaikumar, A. Nayyar, Convolutional neural network based early fire detection, Multimedia Tools and Applications 79 (13) (2020) 9083–9099.
[28] G. Zhang, M. Wang, K. Liu, Forest fire susceptibility modeling using a convolutional neural network for yunnan province of china, International Journal of Disaster Risk Science 10 (3) (2019) 386–403.
[29] N. Dilshad, T. Khan, J. Song, Efficient deep learning framework for fire detection in complex surveillance environment., Comput. Syst. Sci. Eng. 46 (1) (2023) 749–764.
[30] T. Khan, Z. A. Khan, C. Choi, Enhancing real-time fire detection: An effective multi-attention network and a fire benchmark, Neural Computing and Applications 37 (18) (2025) 11693–11707.
[31] Z. Wang, L. Xu, Z. Chen, Ffd-yolo: A modified yolov8 architecture for forest fire detection, Signal, Image and Video Processing 19 (3) (2025) 265.
[32] K. Lv, R. Wu, S. Chen, P. Lan, Cci-yolov8n: Enhanced fire detection with carafe and context-guided modules, in: International Conference on Intelligent Computing, Springer, 2025, pp. 128–140.
[33] S. Li, Q. Yan, P. Liu, An efficient fire detection method based on multi-scale feature extraction, implicit deep supervision and channel attention mechanism, IEEE Transactions on Image Processing 29 (2020) 8467–8475.
[34] F. Yuan, K. Li, C. Wang, Z. Fang, A lightweight network for smoke semantic segmentation, Pattern Recognition 137 (2023) 109289.
[35] H. Yar, T. Hussain, M. Agarwal, Z. A. Khan, S. K. Gupta, S. W. Baik, Optimized dual fire attention network and medium-scale fire classification benchmark, IEEE Transactions on Image Processing 31 (2022) 6331–6343.
[36] H. Yar, Z. A. Khan, T. Hussain, S. W. Baik, A modified vision transformer architecture with scratch learning capabilities for effective fire detection, Expert Systems with Applications 252 (2024) 123935.
[37] W. Wang, L. Zhang, T. Yang, S. Ma, Q. Zhang, P. Shi, F. Ding, Combined serum free light chain predicts prognosis in acute kidney injury following cardiovascular surgery, Renal Failure 44 (1) (2022) 1–10.
[38] S. Majid, F. Alenezi, S. Masood, M. Ahmad, E. S. Gündüz, K. Polat, Attention based cnn model for fire detection and localization in real-world images, Expert Systems with Applications 189 (2022) 116114.
[39] M. Jiang, Y. Zhao, F. Yu, C. Zhou, T. Peng, A self-attention network for smoke detection, Fire safety journal 129 (2022) 103547.
[40] Z. A. Khan, F. U. M. Ullah, H. Yar, W. Ullah, N. Khan, M. J. Kim, S. W. Baik, Optimized cross-module attention network and medium-scale dataset for effective fire detection, Pattern Recognition 161 (2025) 111273.
[41] X. Yu, C.-H. Chen, A robust operators' cognitive workload recognition method based on denoising masked autoencoder, Knowledge-Based Systems 301 (2024) 112370.
[42] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16000–16009.
[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[45] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, S. W. Baik, Convolutional neural networks based fire detection in surveillance videos, Ieee Access 6 (2018) 18174–18183.
[46] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, S. W. Baik, Efficient deep cnn-based fire detection and localization in video surveillance applications, IEEE Transactions on Systems, Man, and Cybernetics: Systems 49 (7) (2018) 1419–1434.
[47] K. Muhammad, S. Khan, M. Elhoseny, S. H. Ahmed, S. W. Baik, Efficient fire detection for uncertain surveillance environment, IEEE Transactions on Industrial Informatics 15 (5) (2019) 3113–3122.
[48] Z. Deng, S. Hu, S. Yin, Y. Wang, A. Basu, I. Cheng, Multi-step implicit adams predictor-corrector network for fire detection, IET Image Processing 16 (9) (2022) 2338–2350.
[49] J. Feng, Y. Sun, Multiscale network based on feature fusion for fire disaster detection in complex scenes, Expert Systems with Applications 240 (2024) 122494.
[50] H. Yar, W. Ullah, Z. A. Khan, S. W. Baik, An effective attention-based cnn model for fire detection in adverse weather conditions, ISPRS Journal of Photogrammetry and Remote Sensing 206 (2023) 335–346.
[51] H. Yar, Z. A. Khan, I. Rida, W. Ullah, M. J. Kim, S. W. Baik, An efficient deep learning architecture for effective fire detection in smart surveillance, Image and Vision Computing 145 (2024) 104989.
[52] S. B. Avula, S. J. Badri, G. Reddy, A novel forest fire detection system using fuzzy entropy optimized thresholding and stn-based cnn, in: 2020 international conference on COMmunication Systems & NETworkS (COMSNETS), IEEE, 2020, pp. 750–755.
[53] T.-H. Chen, P.-H. Wu, Y.-C. Chiou, An early fire-detection method based on image processing, in: 2004 International Conference on Image Processing, 2004. ICIP'04., Vol. 3, IEEE, 2004, pp. 1707–1710.
[54] T. Celik, H. Ozkaramanli, 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.
[55] D. Zhang, S. Han, J. Zhao, Z. Zhang, C. Qu, Y. Ke, 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.

Title Page

Masked Autoencoder-Based Vision Framework for Robust Fire Detection in Complex Environments
Authors:

1. Hanxiang Wang †
Research Assistant
School of Computer Science, Qufu Normal University, Rizhao, China
Email: hanxiang@qfnu.edu.cn

2. Muhammad Fayaz
Research Assistant
Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
Email: muhammadfayaz@sju.ac.kr

3. Awais Ahmad
Research Assitant
Department of Computer Science, Islamia college university, Peshawar 25000, Pakistan
Email: islamian1@gmail.com

4. Yanfen Li *
Research Assistant
School of Computer Science, Qufu Normal University, Rizhao, China
Email: hanxiang@qfnu.edu.cn

5. Tan N. Nguyen †
Assistant Professor
Department of Architectural Engineering, Sejong University, Seoul, 05006, South Korea,
Email: tnnguyen@sejong.ac.kr

Corresponding Author:

Prof: L. Minh Dang *
Professor
The Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam,
Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam,
Email: danglienminh@duytan.edu.vn

Keywords:

- Fire Detection
- Masked Autoencoder
- Vision Transformer
- Self-Supervised Learning
- Feature Reconstruction
- Attention Mechanism

† Hanxiang Wang and Tan N. Nguyen are co-first authors and contributed equally to the work.
* These authors share corresponding authorship

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: