A Dynamic Multimodal Fusion Framework for Forest Fire Detection

Ying Xie[1], Qi Jin[2*], Xiaosong Zhang[3], Xuyang Ding[4]
[1]*College of Computer Science and Artificial Intelligence , Southwest Minzu University* , Chengdu , China, xieying@swun.edu.cn
[2*]*School of Computer Science and Engineering , University of Science and Technology of China*, Chengdu, China, 2455659121@qq.com
[3]*School of Computer Science and Engineering , University of Science and Technology of China*, Chengdu, China, johnsonsxz@uestc.edu.cn
[4] *School of Computer Science and Engineering, University of Science and Technology of China*, Chengdu, China, dxy@uestc.edu.cn

***Abstract* — Accurate fire detection is crucial for forest fire prevention and ecological protection. Existing multimodal models face challenges in inter-modal data fusion and feature extraction, leading to performance degradation under complex backgrounds and environmental disturbances. To address these issues, this paper proposes a dynamic multimodal data fusion framework that integrates RGB images, infrared images, and environmental sensor data. First, environmental sensor data are normalized and visualized into image representations, then spatially aligned with the visual modalities to ensure consistency and fusion feasibility. Next, the multiscale feature extraction and optimization module, together with the global feature modeling module, is employed to capture both local and global features. Channel and spatial attention mechanisms are incorporated to enhance the representation of key fire-related regions, including flames, smoke, and high-temperature areas. Finally, a fusion layer is used for deep joint modeling of multimodal features. Experimental results demonstrate that the proposed method outperforms both unimodal and existing multimodal approaches in overall performance and classifier adaptability.**

Keywords—Forest Fire Detection, Multimodal Fusion, Feature Extraction, Channel Attention Mechanism, Spatial Attention Mechanism

I. INTRODUCTION

Forest fires represent one of the most devastating natural disasters, inflicting severe damage on both ecosystems and human society. Beyond the destruction of property, forest fires significantly threaten biodiversity, degrade air quality, and disrupt the global carbon cycle. Consequently, early detection and real-time monitoring of forest fires are critical for minimizing losses and preserving ecological balance.

Traditional forest fire monitoring methods primarily rely on satellite remote sensing, ground patrols, and fixed observation towers. However, satellite-based monitoring is often hindered by cloud cover and limited temporal resolution, making it difficult to detect early-stage and small-scale fires promptly and accurately. Ground patrols, while flexible, suffer from limited coverage, high labor costs, and reduced efficiency in rugged terrain. On the other hand, fixed observation are constrained by line-of-sight limitations and provide only localized monitoring, while also requiring significant construction and maintenance costs.

With the development of computer vision and sensor technologies, image recognition-based forest fire monitoring has emerged as a prominent research direction. Most existing image recognition-based methods rely on unimodal data sources, such as RGB images, infrared images, and environmental sensor data. However, each modality of data source has inherent limitations. Infrared images are highly susceptible to interference from non-fire heat interference and background radiation, leading to high false-positive rates and insufficient understanding of contextual information. RGB images are significantly affected by lighting and smoke conditions, perform poorly in low-light environments, and cannot detect heat sources without visible flames. Environmental sensor data suffers from limited spatial coverage, complex interpretation, and transmission delays, often resulting in false alarms and incomplete fire condition assessments.

Although recent deep learning-based approaches have enhanced forest fire detection accuracy via feature-level fusion, they still have the following limitations in fully leveraging the complementary advantages of multimodal data: (1) Existing multimodal fusion methods underutilize environmental sensor data, neglecting key parameters such as temperature and humidity, which limits model adaptability in complex scenarios; (2) Structural mismatches between image and non-image modalities lead to poor inter-modal consistency, constraining deep cross-modal fusion; (3) Imbalanced feature representation strategies cause insufficient coordination between local and global features, reducing sensitivity to fine-grained fire cues; (4) The lack of dedicated fire-feature enhancement mechanisms makes models prone to false or missed detections, thereby weakening robustness and generalization in real-world conditions.

To address the above limitations of existing multimodal models, this paper proposes a novel multimodal data fusion framework for forest fire detection. This framework transforms environmental sensor data, such as temperature and humidity data, into image-like representations, enriching the overall feature space, and fuses it with RGB images and infrared images, achieving intuitive fusion of heterogeneous modalities. Through optimized feature selection and interpretability-enhanced visualization operations, this framework not only improves detection performance but also enhances model transparency. The main contributions of this paper are summarized as follows:

- A visualization embedding strategy for environmental sensor data is introduced, which encodes parameters such as temperature and humidity into an image-compatible format

and captures their correlation with fire detection labels through polar coordinate visualization. This strategy not only enables modality alignment between RGB and infrared images but also enhances the model's robustness in complex environments by emphasizing key regions associated with fire risk.

· A multiscale feature extraction and optimization module and a global feature modeling module were designed to enable accurate identification and classification of flammable regions. The multiscale feature extraction and optimization module effectively capture local feature patterns. In contrast, the global feature modeling module is responsible for learning the overall semantic context, providing complementary information to enhance the understanding of flammable regions.

· A multimodal fusion method based on a dual-channel attention mechanism is proposed. By adaptively highlighting key modes and key regions through cross-modal channel attention and spatial saliency attention, more efficient fire feature fusion and recognition can be achieved.

II. RELATED WORKS

This section provides a comprehensive review of existing multimodal forest fire recognition methods.

In recent years, various deep learning models have been proposed, particularly convolutional neural networks (CNNs) and Transformers, leverage end-to-end learning to automatically extract and integrate features from multimodal data, significantly improving detection accuracy and real-time performance. Wang et al. [1] proposed the Fire in Focus algorithm with a boundary enhancement mechanism to improve fire region identification, while Niu et al. [2] enhanced YOLOv5s-Seg through multi-scale feature fusion and attention modules for small fire detection. Mu et al. [3] introduced a superpixel-based graph convolutional network (GCN) to reduce computational cost while maintaining segmentation accuracy, and Abdusalomov et al. [4] developed an improved forest fire detection framework based on Detectron2 with a custom-annotated dataset. Reis et al. [5] utilized drone-based fire imagery and applied transfer learning using InceptionV3 and DenseNet121 to enhance detection performance. Wu et al. [6] combined CNNs and Transformers in a multi-scale architecture, capturing local visual cues through CNNs and global context via Transformers for robust detection. In addition to these studies, Zhou et al. [7] presented MTANet with hierarchical multimodal fusion for efficient RGB-T integration, and Sun et al. [8] designed a multimodal Transformer resilient to data misalignment, improving the reliability of multimodal fire detection. Huang et al. [9] proposed a Deformable Transformer-based model for small-object smoke detection with refined bounding boxes, while Zhu et al. [10] and Mai et al. [11] explored image-text interaction and hybrid contrastive learning frameworks adaptable to fire detection for semantic enhancement. Furthermore, Ghali and Akhloufi [12] reviewed state-of-the-art deep learning approaches for wildfire detection, mapping, and prediction using remote sensing, and Zhang et al. [13] introduced Ship-Fire Net, an improved YOLOv8-based model employing long-range attention and optimized convolution to boost ship fire detection accuracy and efficiency.

Building upon these developments, multimodal fire detection integrates diverse data sources to further enhance accuracy and reliability. Vikram and Sinha [14] combined visual and thermal imagery for forest fire monitoring, while Sharma et al. [15] applied multimodal and federated learning for urban scenarios. Chen et al. [16] utilized drone-based RGB/IR datasets, and Bhamra et al. [17] developed Multimodal SmokeyNet, which fuses satellite imagery, meteorological sensor data, and optical inputs to improve detection accuracy and response time. Zhou et al. [18] proposed MTANet for hierarchical RGB-T fusion, and Li et al. [19] introduced a multimodal graph learning method for complex data integration. Chaoxia et al. [20] proposed a weakly aligned multimodal flame detection framework for firefighting robots, integrating projection and attention-guided modules to mitigate alignment issues and enhance robustness.

Despite these advancements, several limitations remain. Most fusion methods, such as those proposed by Bhamra et al. [17], remain feature-level, with unresolved spatial alignment between image and sensor data. Models employing attention mechanisms, including those by Niu et al. [2] and Chaoxia et al. [20], still suffer from false alarms under complex conditions. Although some studies incorporate attention mechanisms and multiscale feature extraction to improve small fire detection, their robustness in complex, noisy environments remains limited. Model performance also deteriorates significantly when data are misaligned or incomplete. Furthermore, recent approaches that focus on enhancing fusion strategies often neglect the spatial and temporal dependencies between modalities, reducing adaptability in dynamic fire scenarios. Key challenges for future research therefore include achieving precise modality alignment, improving fusion robustness, and fully leveraging cross-modal complementarity to enable reliable, real-time fire detection in complex environments.

III. METHODOLOGY

The overall of the proposed dynamic multimodal fusion framework in this paper is shown in Figure 1, which involves t input and pre-processing, image fusion, and classification prediction.

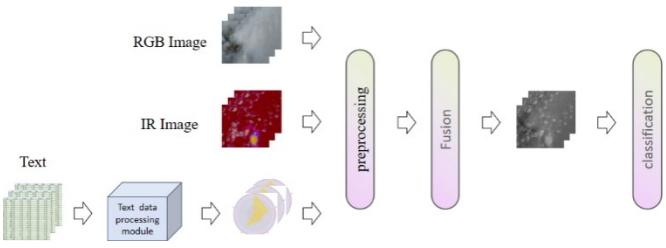

Fig. 1. Dynamic Multimodal Fusion Framework

Environmental sensor data are normalized and visualized into image-like representations aligned with RGB and IR images to ensure multimodal consistency. After pre-processing, to improve detection accuracy, a multi-scale feature extraction and optimization module is proposed to capture local features; a global feature modeling module is proposed to capture overall contextual information; and channel and spatial attention mechanisms are proposed to enhance the representation of flame, smoke, and high-temperature regions.

A. *Pre-processing module*

Environmental sensor data is numerical and requires initial cleaning to remove noise and outliers. It is then normalized to maintain its relative relationships and proportions during the transformation process:

$$x_{i,j} = \frac{s_{i,j} - \mu_j}{\sigma_j} \tag{1}$$

where $s_{i,j} \in D$ denotes the original value of the j-th environmental variable for the i-th sample in dataset D, μ_j and σ_j are the mean and the standard deviation of the j-th variable over the entire dataset D, and $x_{i,j}$ is the normalized value of the j-th variable for the i-th sample.

The normalized data are then converted into image form using visualization techniques. These numerical images are resized to match the spatial dimensions of the other image modalities, ensuring spatial consistency across all modalities. The workflow of text data processing is shown in Figure 2.

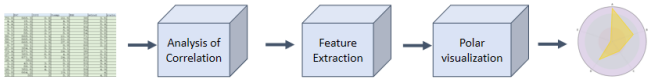

Fig. 2. Text data processing workflow

This study proposes the combustion potential index (CP), which takes into account environmental variables including temperature (T), wind speed (W), small fuel humidity code (FFMC), the concentrations of PM2.5 and PM1.0, and quantifies the potential contribution of these variables to fire risk. The formula of CP index is：

$$CP_i = w_1 \cdot T_i + w_2 \cdot W_i - w_3 \cdot FFMC_i + w_4 \cdot PM1.0_i + w_5 \cdot PM2.5_i \tag{2}$$

$$w_j = \frac{\sigma_j}{\sum_{k=1}^{5} \sigma_k}, j = 1,2,\dots,5 \tag{3}$$

where CP_i represents the combustion potential of the i-th sample corresponding to the normalized environmental variables, σ_j denotes the standard deviation of the j-th environmental variable, representing the variability in the importance of each variable; w_j is the weight of the j-th variable; $w_3 < 0$ indicates that the higher the fuel humidity, the lower the fire potential.

To quantify the relationship between environmental variables and fire occurrence labels, a risk enhancement correlation coefficient based on the Pearson correlation coefficient (PCC) was proposed.

$$r_j' = \frac{\sum_{i=1}^{n} \alpha_j \cdot (x_{i,j} - \bar{x}_j)(y_i - \bar{y}) + \beta \cdot CP_i}{\sqrt{\sum_{i=1}^{n} \alpha_j \cdot (x_{i,j} - \bar{x}_j)^2} \cdot \sqrt{\sum_{i=1}^{n} (\alpha_j \cdot (y_i - \bar{y})^2 + \gamma \cdot CP_i^2)}} \tag{4}$$

where r_j' is the risk-enhanced correlation coefficient between the j-th environmental variable and the occurrence of fire , $x_{i,j}$ and y_i denote the j-th environmental variable and fire labels of the i-th sample, respectively; $\bar{x}_j$ and $\bar{y}$ represent the means of the j-th environmental variable and the file label; and $\alpha_j$represents the importance weight of the j-th variable obtained from the Random Forest feature importance evaluation. β and γ are the weight parameters to regulate the effect of CP_i.

Furthermore, polar coordinate visualization was employed to illustrate the contributions of these selected features to fire risk and their interrelationships with respect to fire occurrence.

The essence of polar visualization is to represent the relative importance and directionality of variables in relation to fire risk through their amplitude and angle . Amplitude represents the comprehensive contribution of variables to fire risk and the Angle represents the directional relationship of each variable with respect to fire risk. The amplitude $\mathcal{R}_j$ and angular position θ_j of the j-th environmental variable in the polar coordinate system are defined as follows:

$$\mathcal{R}_j = \sqrt{\sum_{i=1}^{n} ((x_{i,j} + \lambda \cdot CP_i) \cdot r_j')^2} \tag{5}$$

$$\theta_j = \tan^{-1}\left(\frac{\sum_{i=1}^{n} (x_{i,j} + \lambda \cdot CP_i) \cdot r_j'}{\sum_{k=1,k\neq j}^{m} \sum_{i=1}^{n} (x_{i,k} + \lambda \cdot CP_i) \cdot r_k'}\right) \tag{6}$$

where the risk-enhanced correlation coefficient r_j' is used as the weight of the variable j, and λ is the adjustment coefficient of combustion potential.

Fig. 3 shows an example of the polar coordinate visualization, where each vertex represents a variable contributing to fire risk. The radius $\mathcal{R}_j$ indicates the strength of each variable's influence, with larger radius corresponding to stronger risk contributions and smaller radius indicating weaker impacts. The angle θ_j reflects its directional relationship to fire occurrence, where larger angles indicate that the variable's influence pattern differs more from other variables, while smaller angles suggest a more similar or aligned influence pattern. The grey polygon formed by connecting these points shows the combined influence and interaction of variables: its enclosed area reflects the aggregated contribution of all variables to fire risk, where a larger area represents a higher overall fire risk level, whereas a smaller area suggests lower aggregated risk.

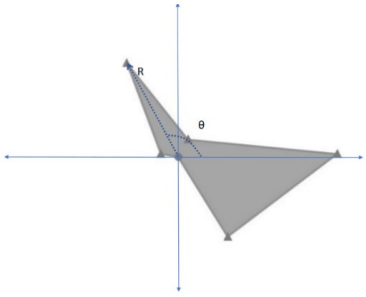

Fig. 3. Example of polar coordinate visualization

The input phase of the model involves receiving paired RGB and infrared (IR) images, along with numerical environmental data that have been converted into image-like modalities. To ensure data consistency and comparability in subsequent processing, the pre-processing module performs dimensional alignment and normalization on all input data. In particular, all

image modalities are resized to a uniform resolution using bilinear interpolation to facilitate pixel-level fusion.

B. *Dynamic fusion module*

Dynamic fusion module aims to efficiently extract and fuse salient features from multiple pre-processed multimodal input images, thereby enhancing the accuracy and reliability of fire detection. The overall workflow of this module includes the encoder, fusion layer and decoder.

1) Encoder

The core function of the encoder is to progressively extract key fire-indicative features from the input multi-channel images, capturing information from local details to global representations. Let the input image be $I \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, H and W are the height and width of the image, respectively.

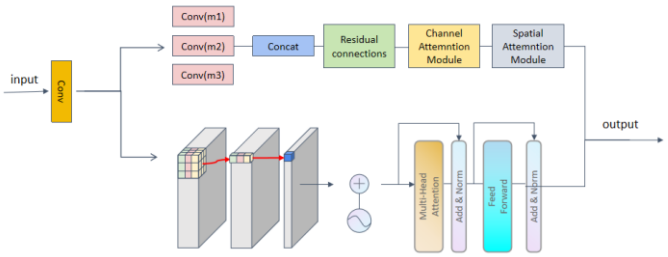

Fig. 4. Encoder structure

a) Multiscale feature extraction and optimization module

The convolution operations are used to extract low-level features in our encoder. The output feature image F_{conv} of image I can be expressed as:

$$F_{conv} = ReLU(W_{conv} * I + b) \tag{7}$$

where W_{conv} is the convolution kernel, b is the offset term. F_{conv} is the low-level feature map extracted by the first convolutional layer of the encoder.

Then, multiple convolution cores are used to extract multi-scale features. Each convolution kernel extracts features under different receptive fields: small convolution kernel is suitable for extracting local details of the flame edge; Medium convolution kernels are suitable for capturing the local texture of smoke diffusion; Large convolution kernels are suitable for modeling the global distribution of fire areas. The convolution operations at the small, medium, and large scales are as follows:

$$F_{scale1} = ReLU(W_{m_1\times m_1} \cdot F_{conv} + b_{m_1\times m_1}) \tag{8}$$

$$F_{scale2} = ReLU(W_{m_2\times m_2} \cdot F_{conv} + b_{m_2\times m_2}) \tag{9}$$

$$F_{scale3} = ReLU(W_{m_3\times m_3} \cdot F_{conv} + b_{m_3\times m_3}) \tag{10}$$

where $F_{scale1}, F_{scale2}, F_{scale3}$ represent the feature map extracted using convolution kernels of sizes m_1, m_2 *and* m_3. $W_{m_1\times m_1}, W_{m_2\times m_2}, W_{m_3\times m_3}$ represent the convolution kernels with sizes m_1, m_2 *and* m_3 . $b_{m_1\times m_1}, b_{m_2\times m_2}, b_{m_3\times m_3}$ represent the bias terms associated with each convolution kernel.

These features are stitched together to form a multi-scale feature map , allowing each layer to incorporate information from all preceding feature scales. The resulting multi-scale feature map is represented as $F_{Multi-scale}$:

$$F_{Multi-scale} = Concat(F_{scale1}, F_{scale2}, F_{scale3}) \tag{11}$$

After multi-scale feature extraction, these concatenated features are further optimized through residual connections to strengthen deep semantic representation and avoid gradient disappearance. Let the output after the residual connection be F_{Res}:

$$F_{Res} = F_{Multi-scale} + \mathcal{F}(F_{Multi-scale}, \{W_i\}) \tag{12}$$

$$\mathcal{F}(F_{Multi-scale}, \{W_i\}) = ReLU(W_2 \cdot ReLU(W_1 \cdot F_{Multi-scale} + b_1) + b_2) \tag{13}$$

where W_1, W_2 and b_1, b_2 are the weights and bias of the first and second convolutional layers, respectively. $\mathcal{F}(F_{Multi-scale}, \{W_i\})$ represents the residual mapping function applied to the multi-scale feature map using a two-layer convolutional block. F_{Res} represents the residual-enhanced feature map obtained by adding $F_{Multi-scale}$ and its residual mapping.

Finally, Multi-scale and deep features are fused using the Add operation:

$$F_{DenseRes} = F_{Multi-scale} + F_{Res} \tag{14}$$

To enhance the perception of critical fire cues, two complementary attention mechanisms are introduced: temperature attention, which emphasizes heat anomalies derived from RGB–IR differences, and smoke attention, which highlights spatial patterns associated with smoke and flame regions. Together, they strengthen the model's sensitivity to temperature and smoke variations.

Temperature attention identifies heat anomalies by contrasting infrared (IR) and visible (RGB) features, and then generates temperature-aware feature map $F_c^{enhanced}$ that highlights high-temperature regions in the feature space:

$$T_{anomaly} = ReLU(F_{IR} - F_{RGB}) \tag{15}$$

$$M_c^{temp} = Sigmoid(W_c \cdot Sigmoid(M_{avg} + M_{max} + GAP(T_{anomaly})) + b_c) \tag{16}$$

$$F_c^{enhanced} = F_{DenseRes} \odot M_c^{temp} \tag{17}$$

where F_{IR} and F_{RGB} denote IR and RGB feature maps, $T_{anomaly}$ is the temperature anomaly map, M_{avg} and M_{max} are global average and maximum pooling result of $F_{DenseRes}$, $GAP(\cdot)$ represents global average pooling, M_c^{temp} is the temperature-aware channel weight, and $\odot$ denotes channel-by-channel weighting.

Smoke attention enhances spatial awareness by integrating edge information and smoke saliency, generating spatial-aware feature map $F_s^{enhanced}$ that highlight smoke and flame regions:

$$S_{smoke} = Sobel(F_{gray}) \tag{18}$$

$$M_s^{smoke} = Sigmoid(W_s * [M_{avgpool}, M_{maxpool}, S_{smoke}]) \tag{19}$$

$$F_s^{enhanced} = F_c^{enhanced} \otimes M_s^{smoke} \tag{20}$$

where F_{grey} is the greyscale feature , S_{smoke} is the smoke saliency map obtained from F_{grey} using the Sobel operator. $M_{avgpool}$ and $M_{maxpool}$ are global pooling and max pooling maps of $F_c^{enhanced}$, M_s^{smoke} is the smoke-based spatial weight map, * denotes a convolution operation, ⊗ denotes pixel-by-pixel weighting. By sequentially applying the temperature-aware channel enhancement and the smoke-aware spatial enhancement, the final fire-scene enhancement feature $F_{Enhanced}$ is obtained, which is represented by $F_s^{enhanced}$.

b) Global feature modeling module

The global feature modeling module is designed to capture global correlations and long-range dependencies among multimodal input features, complementing the local details extracted by the multiscale feature extraction and optimization module. The input feature F_{in} is first divided into multiple fixed-size patches, each representing a low-dimensional feature of a local region. These patches are then linearly projected into an embedding space to preserve their spatial order and form the input to the attention mechanism.

To model the global relationship among patches, the module employs Multi-Head Self-Attention. By applying linear transformations to the embedded features, the query (Q), key (K), and value (V) matrices are obtained, and the attention weights are calculated as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)\cdot \mathrm{V} \qquad (21)$$

where d_k is the dimensionality of each attention head , which is used to scale the inner product result to avoid gradient explosion.

The multi-head mechanism allows the model to learn diverse subspace representations, and the outputs of all heads are concatenated and linearly transformed to form the MHSA output. This output is further refined through a Feed Forward Network (FFN) with GELU activation and residual connections, followed by layer normalization to ensure stable feature distribution. The resulting feature F_{GFM} is reconstructed to match the original spatial structure, representing the globally enhanced feature map.

c) The output

The multiscale feature extraction and optimization module focuses on local features, enhancing detailed spatial information through multiple receptive fields and feature refinement, while the global feature modeling module focuses on capturing contextual and long-range dependencies. We concatenate the outputs of these two modules, the local features $F_{Enhanced}$ and the global features F_{GFM} , along the channel dimension, as follows:

$$F_{Concat} = Concat(F_{Enhanced}, F_{GFM}) \qquad (22)$$

This concatenation operation combines fine-grained local details with global contextual information, enabling a more comprehensive feature representation of the fire scene and providing stronger multimodal input for subsequent fusion and decoding.

2) Fusion Layer

The fusion layer combines the multimodal features of the fire scene and designs an improved composite attention mechanism to effectively capture the spatial distribution of the fire region and the correlation between channels.

a) Feature space reconstruction module

In the fusion stage, z_1 and z_2 denote the modality-specific concatenated feature maps derived from F_{Concat} for the two input modalities in our system . Each input feature tensor z_1 and z_2 is subsequently fed into the feature space reconstruction module, which operates in two steps. (1) Feature compression and expansion: 1×1 convolution is used to perform channel dimensionality reduction on the input features to reduce redundant information and generate preliminary feature representations. (2) Biaxial feature modelling : During the reconstruction process, features are alternately aggregated along the height and width axes, implemented through $SpatialReconstruct(\cdot)$, to capture the global feature correlations in 2D space. This process produces the reconstructed feature representations $f_1$and f_2.

$$f_1 = SpatialReconstruct\left(z_1\right)$$
$$f_2 = SpatialReconstruct\left(z_2\right) \qquad (23)$$

b) Fire feature channel enhancement module

To emphasize key fire-related channel information (e.g., heat, smoke, and flame) during the fusion phase, we introduces a dynamic weighting strategy.

First, Global Average Pooling (GAP) and Global Max Pooling (GMP) are applied to each channel of the input feature map $f \in \mathbb{R}^{C\times H\times W}$, generating two contextual descriptors that represent the global and most activated regions of fire, respectively. Then, the pooled context vectors are combined and passed through a shared Multi-Layer Perceptron (MLP) to produce dynamic weight coefficients a_1 and a_2 , which adaptively adjust channel responses. The channel-enhanced version of f_1 and f_2 are denoted as follows:

$$\widetilde{f_1} = a_1 \cdot f_1$$
$$\widetilde{f_2} = a_2 \cdot f_2 \qquad (24)$$

c) The Output

The enhanced feature maps are then fused through pixel-wise addition to produce the final fused feature F_{fusion}. This fused representation integrates information from multiple modalities, capturing both the global spatial dependencies of fire regions and the dynamic interactions across channels, thereby yielding stronger fire identification capabilities.

$$F_{fusion} = \widetilde{f_1} + \widetilde{f_2} \qquad (25)$$

3) Decoder

The main function of the decoder is to reconstruct the fused feature map into an output image Y. We realize this process by inverse convolution and up-sampling. The mathematical expression of the decoder is:

$$F_{out} = Dec\left(F_{fusion}\right) = W_{deconv} * F_{fusion} + b_{deconv}$$
$$Y = \mathrm{Sigmoid}(F_{out}) \qquad (26)$$

where W_{deconv} represents the deconvolution (transpose convolution) kernel of the decoder, and b_{deconv} represents the corresponding bias term. F_{out} represents the output feature map generated by the decoder before the final activation.

Because the transposed convolution and up-sampling operations progressively restore spatial resolution, while the learnable deconvolution kernels map high-level fusion semantics back to pixel-level representations, this structure ensures that the decoder can effectively transform fused features into meaningful output.

IV. EXPERIMENTS

A. The Dataset

This study utilizes multiple datasets, integrating environmental sensor data and image sources to ensure a comprehensive analysis. The Montesano Natural Park dataset [21] provides key meteorological and fire-related variables; the Kaggle dataset [22] includes labeled RGB images of fire and non-fire scenarios; the FLAME2 dataset [23] offers paired RGB–IR aerial imagery; and an additional smoke detection dataset[24] contains temperature, humidity, particulate matter, and fire alarm indicators.

The Montesano Natural Park [21] includes multiple parameters relevant to forest fire scenarios, all of which are critical for assessing fire risk. In our experiments, several irrelevant features—such as month, X and Y coordinates, and date—were removed, and the final dataset comprised the following nine features: relative humidity (RH), wind speed (Wind), region area (Area), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), rainfall (Rain), temperature (Temp), and Fine Fuel Moisture Code (FFMC). The details of each feature are summarized in Table Ⅰ.

TABLE I. THE RANGE OF EACH FEATURE

Features	Span
RH	15.0–100
Wind	0.40-9.40
Area	0.00–1090.84
DMC	1.1–291.3
DC	7.9–860.6
ISI	0.0–56.10
Rain	0.0–6.4
Temp	2.2–33.30
FFMC	18.7–96.20

All images in the Kaggle dataset [22] were categorized into two classes: “fire” and “non-fire.” For the purposes of this study, a total of 517 images were selected, comprising 270 fire images and 247 non-fire images. Sample images from each category are illustrated in Fig. 5 and 6.

Fig. 5. Sample of Fire images

Fig. 6. Sample of Non-Fire images

The FLAME2 dataset [23] comprises 53,451 paired dual-view frames, each containing one RGB image and a corresponding infrared (IR) image, spatially aligned side by side to facilitate comparative analysis. As illustrated in Figs. 7 and 8, representative examples of “fire” and “non-fire” RGB–IR image pairs are presented.

Fig. 7. Sample of Fire RGB and IR images

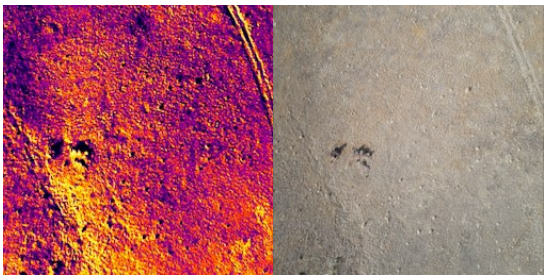

Fig. 8. Sample of Non -Fire RGB and IR images

Additionally, the study incorporated data collected from a smoke detection system [24], which included several key variables: ambient temperature (°C), relative humidity (%), concentrations of particulate matter with diameters of 1.0 μm (PM1.0) and 2.5 μm (PM2.5), and a binary fire alarm indicator representing whether a fire alarm was triggered (1 for alarm, 0 for no alarm).

These datasets were partitioned into two mutually exclusive subsets, with 50% allocated for training and the remaining 50% reserved for testing. This strict separation—analogous to evaluation standards in the medical domain—was designed to prevent data leakage and overfitting, thereby providing a more accurate assessment of the model’s performance on previously unseen data.

In this study, the performance of the model is evaluated through four key metrics, including accuracy, precision, recall, and F1 score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{28}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{29}$$

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{30}$$

where *TP* denotes the number of true positives, *TN* the true negatives, *FP* the false positives, and *FN* the false negatives. A higher accuracy value indicates better overall model performance.

B. Performance Evaluation

To comprehensively validate the effectiveness of the proposed multimodal data fusion approach, the experiments were conducted and analyzed from three distinct perspectives:

- Performance comparison between unimodal and multimodal models;
- Performance comparison between multimodal models;
- Generalizability Evaluation.

The results corresponding to these three aspects are summarized in Tables 3, 4, and 5, respectively. Detailed analyses are provided below.

1) Performance comparison between unimodal and multimodal models

TABLE II. UNIMODAL VS. MULTIMODAL MODELS PERFORMANCE

Mode	Accuracy	Precision	Recall	F1
RGB Only[23]	95.33	95.86	95.12	95.39
Thermal Only[23]	92.15	92.19	92.79	92.3
Proposed	**99.65**	**99.75**	**99.63**	**99.69**
RGB Only[22]	79.46	79.68	79.39	79.39
Text Only[24]	89.96	91.72	89.94	89.83
Proposed	**100**	**100**	**100**	**100**
RGB Only[22]	79.46	79.68	79.39	79.39
Text Only[21]	52.12	57.20	51.96	41.35
Proposed	**95.87**	**96.15**	**95.90**	**95.86**

As reported in Table 3, comprehensive experiments were performed under both unimodal and multimodal input conditions. For the image modality, when RGB and infrared (IR) images were used independently, the classification accuracies reached 95.33% and 92.15%, respectively. In contrast, the application of the proposed multimodal fusion strategy—combining RGB and IR data—yielded a significant accuracy improvement, achieving 99.65%. This enhancement underscores the complementary nature of the two modalities: RGB images offer rich visual context, while IR images provide thermal cues that remain informative in low-light or smoke-obscured conditions.

A similar trend was observed in the fusion of text and sensor data. When assessed individually, the unimodal classifiers based on textual descriptions and sensor readings achieved accuracies of 79.46% and 89.96%, respectively. However, when these modalities were integrated using the proposed multimodal framework—which simultaneously captures semantic features from text and temporal dynamics from sensor streams—the classification accuracy reached 100%. This result highlights the model's capability in effectively learning cross-modal dependencies, achieving robust performance across complex and dynamic fire scenarios.

The comparative results clearly demonstrate that unimodal approaches are inherently limited by single-source information constraint, which constrains their robustness and generalizability across diverse fire conditions. In contrast, the proposed multimodal fusion method effectively integrates complementary features from multiple sources, significantly improving classification accuracy. These findings validate the superior representational capacity and adaptability of the proposed approach for forest fire detection.

2) Performance comparison between multimodal models

TABLE III. COMPARISON BETWEEN MULTIMODAL MODELS

Dataset	Metrics	Result from References		Result from our study
RGB Images+ IR[23]	Accuracy	Chen [16]	98.87	**99.65**
	Precision		98.55	**99.75**
	Recall		99.22	**99.63**
	F1 Score		98.86	**99.69**
fire and non-fire images [22]+ Laboratory Fire text[24]	Accuracy	Sharma [15]	98.46	**100**
	Precision		99.15	**100**
	Recall		97.48	**100**
	F1 Score		98.31	**100**
fire and non-fire im ages [22]+ Montesano's natural park text [21]	Accuracy	Vikram [14]	85.12	**95.87**
	Precision		78.87	**96.15**
	Recall		94.92	**95.90**
	F1 Score		86.15	**95.86**

Comparative experiments were performed on three multimodal benchmark datasets against representative multimodal approaches. Since most existing multimodal fire-detection approaches are designed for a single modality pair (e.g., image–image or image–text) rather than supporting multiple modality combinations within a unified framework, representative algorithms were selected to ensure fair comparison. Given the limited availability of multimodal forest fire datasets, three suitable datasets were deliberately chosen to validate the proposed method. The experimental results, summarized in Table 4, include two primary configurations: (1) multimodal fusion of RGB and infrared (IR) images, and (2) multimodal fusion of images and sensor data.

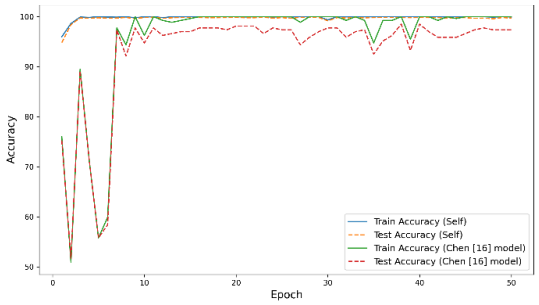

Fig. 9. Training and test accuracy curve of Proposed model and Chen[16] model

Using the RGB + IR [23] dataset, the proposed method achieved a classification accuracy of 99.65%, outperforming the 98.87% accuracy reported by Chen et al. [16], as illustrated in Fig. 9. our model independently extracts features from RGB and IR inputs and adopts a more comprehensive fusion strategy to exploit complementary information between the two modalities. This results in improved robustness and higher detection accuracy.

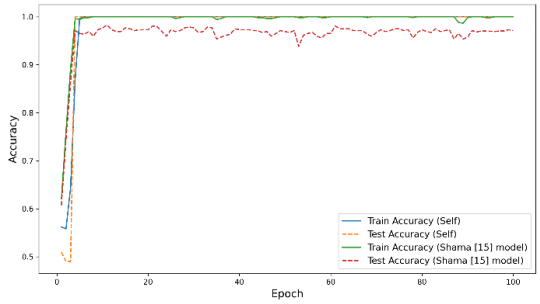

Fig. 10. Training and test accuracy curve of Proposed model and Shama[15] model

On the fire and non-fire images [22]+ Laboratory Fire text[24] , our method achieved a perfect classification accuracy of 100%, clearly outperforming the 98.46% accuracy reported by Sharma et al. [11], as shown in Fig. 10. Compared with our approach, Sharma's method shows limitations in handling heterogeneous modalities. Their framework relies on independently extracted features that are later aggregated, which weakens cross-modal interaction and often leads to incomplete alignment between visual and textual (or sensor-derived) information. Such loosely coupled fusion restricts the model's ability to capture fine-grained semantic correspondence across modalities.In contrast, our method explicitly addresses this weakness by embedding textual descriptions into a unified semantic space and integrating them with image-derived features through a tightly coupled multimodal fusion mechanism. This design facilitates stronger cross-modal correlation learning and avoids the modality inconsistency issues present in Sharma's approach. Consequently, the proposed method can more effectively exploit the complementary nature of image and textual information, leading to superior classification performance.

On the fire and non-fire images [22]+ Montesano's natural park text [21] dataset, our model achieved a classification accuracy of 95.87%, significantly outperforming the 85.12% obtained by the method of Vikram and Sinha [14], as illustrated in Fig. 11. Vikram and Sinha proposed a multimodal fire detection framework that combined image and sensor data using a neural fuzzy classification model (NFCM) to assess fire activity levels (e.g., high, medium, low). Although their method considered multimodal fusion, it did not emphasize multi-scale feature extraction, which may have led to the loss of crucial details when handling fires with varying scale characteristics. In contrast, our method incorporates a multi-scale convolutional network capable of extracting both local features (such as flame and smoke edges) and global features (such as the spatial extent of the fire) across different scales. This allows the model to more precisely capture key fire-related regions, significantly improving fine-grained recognition in complex environments. Overall, the proposed method more effectively leverages the complementary strengths of the two modalities.

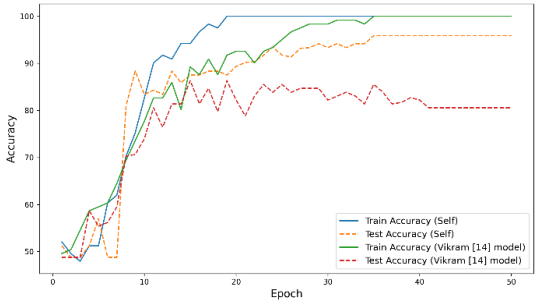

Fig. 11. Training and test accuracy curve of Proposed model and Vikram [14] model

3) Generalizability evaluation

To evaluate the generalizability of the proposed multimodal data fusion method, experiments were conducted by inputting the fused multimodal features into three different classifiers—CNN (used as the baseline in this study), ResNet, and LeNet—for performance comparison. The experimental results are summarized in Table 5.

TABLE IV. COMPARISON ACROSS DIFFERENT CLASSIFIERS USING PROPOSED MODEL.

Model	Dataset	Accuracy	Precision	Recall	F1
CNN	254p RGB Images+254p IR[23]	99.65	99.75	99.63	99.69
Resnet		99.75	99.81	99.69	99.75
Lenet		99.29	99.14	99.50	99.31
CNN	fire and non-fire images [22]+ Laboratory Fire text[24]	100	100	100	100
Resnet		99.23	99.27	99.19	99.23
Lenet		100	100	100	100
CNN	fire and non-fire images [22]+ Montesano's natural park text [21]	95.87	96.15	95.90	95.86
Resnet		95.04	95.83	94.55	94.94
Lenet		94.21	94.38	94.33	94.21

On the 254p RGB Images+254p IR [23], all three classifiers achieved near-saturation performance, with classification accuracies of 99.65% for CNN, 99.75% for ResNet, and 99.29% for LeNet. Among them, ResNet exhibited the strongest ability to extract deep features, while CNN achieved comparable results, only marginally lower. These findings suggest that the effectiveness of the fused features is not highly sensitive to the choice of classifier, indicating strong adaptability and robustness of the proposed fusion method.

On the fire and non-fire images [22] + Laboratory Fire text[24], both CNN and LeNet achieved a perfect classification accuracy of 100%, while ResNet slightly underperformed at 99.23%. Given the high discriminative power of the fused multimodal features generated by our method, even the lightweight LeNet model was capable of achieving performance on par with CNN. This demonstrates the high quality and separability of the learned feature representations.

On the fire and non-fire images [22]+ Montesano's natural park text [21], CNN outperformed the other classifiers, achieving a classification accuracy of 95.87%, followed by ResNet at 95.04% and LeNet at 94.21%. These results indicate that CNN has stronger generalization capabilities in complex

environments and is better suited for handling the variability present in the fused features.

Overall, the observed performance differences across classifiers were relatively minor, and all models exhibited strong classification capabilities. These results confirm that the fused features produced by the proposed multimodal fusion method possess high separability and stability, making them widely applicable across different classifier architectures.

V. CONCLUSION

This study presents a dynamic multimodal fusion framework for forest fire detection, integrating RGB images, infrared (IR) images, and numerical sensor data. Environmental sensor readings are first normalized and transformed into image-like representations, ensuring spatial alignment with the RGB and IR modalities and facilitating deep multimodal data fusion. The framework incorporates a multiscale feature extraction and optimization module to capture local features, alongside a global feature modeling module to capture overarching contextual information. Additionally, channel and spatial attention mechanisms are employed to enhance the representation of flames, smoke, and high-temperature regions, improving detection accuracy. Experiments show that the proposed method outperforms unimodal and existing multimodal approaches, demonstrating superior accuracy, robustness under complex backgrounds, and generalizability across different classifiers and scenarios.

Future work will aim to optimize computational efficiency and explore richer data sources to develop a more intelligent and comprehensive fire detection and monitoring framework.

ACKNOWLEDGMENT

Supported by the Fundamental Research Funds for the Central Universities, Southwest Minzu University (ZYN2025005), Sichuan Science and Technology Program (2024NSFTD0031).

REFERENCES

[1] Wang, G., Wang, F., Zhou, H., & Lin, H. (2024). Fire in Focus: Advancing wildfire image segmentation by focusing on fire edges. Forests, 15(1), 217. https://doi.org/10.3390/f15010217

[2] Niu, K., Wang, C., Xu, J., Yang, C., Zhou, X., & Yang, X. (2023). An improved YOLOv5s-seg detection and segmentation model for the accurate identification of forest fires based on UAV infrared image. Remote Sensing, 15(19), 4694. https://doi.org/10.3390/rs15194694

[3] Mu, Y., Ou, L., Chen, W., Liu, T., & Gao, D. (2024). Superpixel-based graph convolutional network for UAV forest fire image segmentation. Drones, 8(4), 142. https://doi.org/10.3390/drones8040142

[4] Abdusalomov, Akmalbek Bobomirzaevich, Bappy MD Siful Islam, Rashid Nasimov, Mukhriddin Mukhiddinov, and Taeg Keun Whangbo. "An Improved Forest Fire Detection Method Based on the Detectron2 Model and a Deep Learning Approach." Sensors 23, no. 3 (2023): 1512. https://doi.org/10.3390/s23031512.

[5] Reis, Hatice Catal, and Veysel Turk. "Detection of Forest Fire Using Deep Convolutional Neural Networks with Transfer Learning Approach." Applied Soft Computing 143 (2023): 110362. https://doi.org/10.1016/j.asoc.2023.110362.

[6] Wu, S., B. Sheng, G. Fu, et al. "Multiscale Fire Image Detection Method Based on CNN and Transformer." Multimedia Tools and Applications 83 (2024): 49787-49811. https://doi.org/10.1007/s11042-023-17482-4.

[7] Zhou, W., Dong, S., Lei, J., & Yu, L. (2023). MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding. IEEE Transactions on Intelligent Vehicles, 8(1), 48-58. https://doi.org/10.1109/TIV.2022.3164899

[8] Sun, L., Lian, Z., Liu, B., & Tao, J. (2024). Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Transactions on Affective Computing, 15(1), 309-325. https://doi.org/10.1109/TAFFC.2023.3274829

[9] Huang, Jingwen, Jiashun Zhou, Huizhou Yang, Yunfei Liu, and Han Liu. "A Small-Target Forest Fire Smoke Detection Model Based on Deformable Transformer for End-to-End Object Detection." Forests 14, no. 1 (2023): 162. https://doi.org/10.3390/f14010162

[10] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., & Qian, J. (2023). Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia, 25, 3375-3385. https://doi.org/10.1109/TMM.2022.3160060

[11] Mai, S., Zeng, Y., Zheng, S., & Hu, H. (2023). Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 14(3), 2276-2289. https://doi.org/10.1109/TAFFC.2022.3172360

[12] Ghali, Rafik, and Moulay A. Akhloufi. "Deep Learning Approaches for Wildland Fires Using Satellite Remote Sensing Data: Detection, Mapping, and Prediction." Fire 6, no. 5 (2023): 192. https://doi.org/10.3390/fire6050192.

[13] Zhang, Ziyang, Lingye Tan, and Robert Lee Kong Tiong. "Ship-Fire Net: An Improved YOLOv8 Algorithm for Ship Fire Detection." Sensors 24, no. 3 (2024): 727. https://doi.org/10.3390/s24030727.

[14] Vikram, R., & Sinha, D. (2023). A multimodal framework for forest fire detection and monitoring. Multimedia Tools and Applications, 82, 9819-9842. https://doi.org/10.1007/s11042-022-13043-3

[15] Sharma, A., Kumar, R., Kansal, I., Popli, R., Khullar, V., Verma, J., & Kumar, S. (2024). Fire detection in urban areas using multimodal data and federated learning. Fire, 7(4), 104. https://doi.org/10.3390/fire7040104

[16] Chen, X., et al. (2022). Wildland fire detection and monitoring using a drone-collected RGB/IR image dataset. 2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1-4. https://doi.org/10.1109/AIPR57179.2022.10092208

[17] Bhamra, Jaspreet Kaur, Shreyas Anantha Ramaprasad, Siddhant Baldota, Shane Luna, Eugene Zen, Ravi Ramachandra, Harrison Kim, Chris Schmidt, Chris Arends, Jessica Block, et al. "Multimodal Wildland Fire Smoke Detection." Remote Sensing 15, no. 11 (2023): 2790. https://doi.org/10.3390/rs15112790.

[18] Zhou, H., Yu, Y., Wang, C., et al. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering, 7, 743-755. https://doi.org/10.1038/s41551-023-01045-x

[19] Li, M., Zhuang, X., Bai, L., & Ding, W. (2024). Multimodal graph learning based on 3D Haar semi-tight framelet for student engagement prediction. Information Fusion, 105, 102224. https://doi.org/10.1016/j.inffus.2024.102224

[20] Chaoxia, C., W. Shang, F. Zhang, and S. Cong. "Weakly Aligned Multimodal Flame Detection for Fire-Fighting Robots." IEEE Transactions on Industrial Informatics 19, no. 3 (2023): 2866-2875. https://doi.org/10.1109/TII.2022.3158668.

[21] Forest fire dataset (2019) retrieved from http://www3.dsi.uminho.pt/pcortez/forestfires/," 2019

[22] Forest fire image dataset retrived from. https://www.kaggle.com/datasets/phylake1337/fire-dataset.

[23] 254p RGB Images and 254p Thermal Dataset retrived from https://ieee-dataport.org/open-access/flame-2-fire-detection-and-modeling-aerial-multi-spectral-image-dataset

[24] Laboratory Fire Dataset retrived from https://data.mendeley.com/datasets/f3mjnbm9b3/1