Engineering Applications of Artificial Intelligence 152 (2025) 110848

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

From smoke to fire: A forest fire early warning and risk assessment model fusing multimodal data

Peixian Jin [a], Pengle Cheng [a,*], Xiaodong Liu [b,**], Ying Huang [c]

[a] School of Technology, Beijing Forestry University, Beijing, China
[b] School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China
[c] Department of Civil, Construction, and Environmental Engineering, North Dakota State University, ND, USA

ARTICLE INFO

Keywords:
Multimodal smoke risk estimation network
Deep learning
Forest fire monitoring
Multimodal data fusion
Smoke detection

ABSTRACT

Experienced forest firefighters can integrate knowledge of smoke patterns and risk factors to assess fire risk and anticipate potential fire spread and outbreaks in complex, variable environments. This study simulates the previously mentioned monitoring process and presents the Multimodal Smoke Risk Estimation Network (MM-SRENet), an innovative multimodal fusion model. This model uniquely integrates the Multi-Scale Dilation Block and Repulsion Loss into a lightweight and efficient target detection system to accurately identify smoke's presence. Furthermore, an advanced backbone based on star operation extracts the scene characteristics associated with smoke and merges them with various fire risk factors. The objective is to simulate fire risk assessments in smoke scenarios and to reduce the misallocation of fire resources resulting from false alarms and missed alerts. The proposed model was trained and validated on a multimodal dataset comprising multiple backgrounds. It successfully identified smoke features and fire potential risks in different scenarios, achieving a prediction accuracy of 93.06 %. Fusing smoke images with fire risk data resulted in an 18.75 % improvement in recognition accuracy compared to the single modal model. This work bridges the gap between multimodal data fusion and forest fire risk monitoring, providing a new direction for future intelligent forest fire prevention and control practices.

1. Introduction

Forest fires are regarded as a particularly pernicious form of natural disaster, characterized by their sudden onset, extensive damage, and inherent difficulty in their management. They are widely acknowledged as the most significant threat to forest ecosystems and the most pressing challenge currently facing the forestry sector. In conjunction with the intensification of global climate change (Seidl et al., 2017), the frequency of forest fires and their destructive effects are increasing. These have been identified by the international community, including the United Nations, as one of the world's eight major natural disasters and as a major public safety emergency. Forest fire prevention is a public welfare system project spearheaded by the government. With the advent of technology, modern forest fire prevention has gradually incorporated advanced scientific and technological means, including near-earth monitoring (Coban and Bereket, 2020), satellite aerial monitoring (Yuan et al., 2017), and wireless sensor network technology (Vikram et al., 2020), among others, to enhance the early identification of fires and the capacity for rapid response. Smoke is an indicator of the early stages of a forest fire (Long et al., 2022), and its detection can facilitate the early identification of such fires. The combustion products of forest fires are more easily observed in smoke, and visual data are more readily discernible. Therefore, using visual images of the forest environment can enhance the efficiency of smoke detection, thereby improving the early identification of fires.

Despite the considerable advancements in forest fire smoke detection technologies (Chaturvedi et al., 2022), there remain significant limitations inherent to these systems. The presence of smoke does not necessarily indicate the occurrence of a fire. For instance, vapors or mists in a moist environment following precipitation may be misidentified as fire smoke, despite the diminished risk of combustion after a rainfall. Additionally, regular agricultural burning or open burning of waste can impede the system's functionality, frequently resulting in false alarms and ineffectual operation of the forest fire monitoring system. Smoke is

* Corresponding author.
** Corresponding author.
E-mail addresses: jpxschool@bjfu.edu.cn (P. Jin), chengpengle@bjfu.edu.cn (P. Cheng), xd_liu@bjfu.edu.cn (X. Liu), ying.huang@ndsu.edu (Y. Huang).

https://doi.org/10.1016/j.engappai.2025.110848
Received 18 November 2024; Received in revised form 6 March 2025; Accepted 8 April 2025
Available online 11 April 2025

merely one indicator of elevated fire risk; the underlying threat often necessitates further data to accurately ascertain its nature. It is evident that the development of more sophisticated warning technologies is essential for the prevention of false alarms and the optimal functioning of forest fire monitoring systems.

As we know, data sources in forest environments are often diverse (Chinembiri et al., 2024), full of complex correlations and spatial-temporal variations among them, and it is difficult to fully reflect the real risk of fire by visual images alone. To overcome this problem, we try to introduce multimodal information (da Silva et al., 2021) to strengthen the judgment of the model, combined with smoke to better understand the reasonableness of fires occurring in the current environment.

In the field of forest fire prediction, factors such as meteorology, topography, vegetation, combustibles, and human activities are the main driving factors affecting the occurrence of forest fires and the prediction accuracy of models (Abid, 2021), which interact with each other to determine the probability of fire occurrence and development path. In order to reflect the fire risk status of forests more comprehensively, the above parts were selected to be introduced as core fire risk factors and fused with smoke images to identify and detect them, providing more accurate analysis for forest fire early warning, reflecting the practical significance of this study to improve fire risk identification capability in complex environments.

In this study, the MM-SRENet framework is used to implement the fusion of smoke images with fire risk factor data to identify the real fire risk of smoke scenarios in forests. The experimental dataset used for training consists of 3352 image-data pairs of samples built from fire and suspected fire tests with multiple scenarios. Finally, the performance capability of our model is evaluated in terms of accuracy and applicability.

The important contributions of this paper are as follows.

1. To address the challenges of variable shapes and blurred details in smoke detection, the lightweight YOLOv8n model is employed as the smoke detection framework. The Multi-Scale Dilation Block (MSDBlock), which enhances multi-scale perception and detail-capturing capabilities, is introduced into the C2f module of the backbone. Additionally, an extra detection head is added to the original detection structure, improving the model's adaptive detection ability for anomalous smoke shapes.
2. The visual blurring and occlusion effects in forest scenes face similar challenges as crowd detection. We introduce Repulsion Loss into the smoke detection field for the first time, integrated into the loss function, the model can optimize both the attraction of smoke targets and the repulsion of surrounding occluder regions during the training process, to ensure the robust expressiveness of the detector in complex scenes.
3. The proposed MM-SRENet framework integrates detection and modal fusion to support end-to-end output of fire assessment risk. After capturing the presence of smoke, the efficient compact backbone network based on star operation is used to further obtain the smoke scene information, and the fusion of fire risk factors to assist the judgment improves the model comprehension and reliability of the results, which is a key step in realizing automated forest fire assessment.
4. We conducted on-site fieldwork to collect real forest scenes with varying fire risk levels, constructing an experimental dataset comprising 3352 image-data pairs. This self-built dataset encompasses diverse and realistic samples of smoke scenarios under different lighting conditions, such as day and night, as well as various environmental conditions, including drizzle, haze, snowfall, wetness, and dryness. It provides diverse and authentic samples for model training and evaluation.

2. Related work

Smoke is typically the initial indicator of a fire, making it a crucial element in the early detection of such incidents. However, the detection of smoke is more challenging than that of flames (Prema et al., 2022). With the advancement of deep learning and the enhancement of data availability, deep learning-based smoke detection techniques are broadly classified into three categories (Chaturvedi et al., 2022): smoke classification, smoke segmentation, and smoke region localization. Considering that in most smoke detection cases, the goal is to identify the presence of smoke and locate its region rather than performing precise pixel-level segmentation (Yuan et al., 2023), smoke classification and smoke region localization have become the predominant tasks in smoke detection applications. Initial studies relied heavily on the powerful local feature extraction capability of CNN networks to shine in smoke classification and localization tasks. de Venancio et al. (2023) proposed an automatic fire detection method that combines spatial (visual) and temporal patterns in two stages: (i) detecting potential fire events using a CNN for spatial processing, and (ii) analyzing the dynamics of these events over time. Li et al. (2024) presented an early unmanned aerial vehicle (UAV) imagery wildfire smoke detection system using a modified YOLOv7-MS. The use of CNN is key to effectively detecting forest fire incidents based on visual patterns. The later Transformer series of vision algorithms (Jing et al., 2023; Wang et al., 2024) show advantages in capturing global features through the self-attention mechanism, but their high computational complexity and large-scale data requirements are not yet as efficient as CNN in smoke detection tasks with high real-time requirements. There are also those that combine the two, such as Yin et al. (2024) who use a Transformer network and a YOLOv5-based smoke detector to recognize smoke at multiple scales. Chen et al. (2022) proposed the transformer-enhanced convolutional network (TECN) to classify Remote Sensing images containing smoke. The hybrid TECN model exploits the advantages of the CNN and transformer techniques at the same time.

Multimodal research has become a hot topic in the field of Artificial Intelligence over the past few decades. Fire risk identification using multimodal data consists of two main approaches: multimodal image fusion and multisource modal fusion. Rui et al. (2023) noticed the impact of modality-specific features on fire recognition and proposed a new wildfire recognition framework combining Red-Green-Blue (RGB) and Thermal Infrared (TIR) images, which can adaptively learn modality-specific and shared features. Zhao et al. (2024) proposed a Transformer-based Universal Fusion (TUFusion) algorithm with multi-domain fusion capability. TUFusion combines a composite attention fusion strategy that can integrate global and local information. The risk of fire can be observed and inferred from multi-source modal information in the environment, such as flame and smoke images, temperature status, and smoke composition. Previously, multi-sensor fusion techniques have been widely used for fire risk sensing in indoor fire detection (Baek et al., 2021; Shaharuddin et al., 2023; Wu et al., 2021). However, the forest is vast and complex, and the traditional multi-sensor fusion technique is no longer applicable in forest fire detection. Bhamra et al. (2023) proposed SmokeyNet, a deep learning model that integrates multiple data sources and fuses satellite detection, weather sensors, and optical image data for detecting wildfire smoke using spatio-temporal information. Phan et al. (2020) used a multiscale deep neural network model that combines satellite imagery and weather data to detect and localize forest fires at the pixel level and showed that the weather data spatio-temporal alignment could significantly enhance the accuracy of fire detection.

In general, forest fire is a complex process, and smoke alone is not enough to warn of fire. Previous studies have mostly focused on improving the performance of smoke detection (Cao et al., 2019; Hu et al., 2022; Sathishkumar et al., 2023), but neglected the fact that the appearance of smoke does not always represent fire risk. In contrast, our proposed MM-SRENet model, by combining multiple fire risk factor data

with a deep learning framework, achieves both adaptive detection of early forest smoke and integrates forest-oriented multimodal fire risk recognition capabilities, enabling the early warning system to more scientifically simulate and predict fire scenarios similar to the judgmental ability of fire experts.

3. Methodology

3.1. Data collection

3.1.1. Visual data of smoke

Since forest fire smoke collection is a special task involving natural disasters, smoke samples from fires account for a much smaller proportion of the dataset than images of normal forests and are prone to the problem of category imbalance. Previously, synthetic smoke datasets (Yuan et al., 2019) were also used to train the network, but synthetic smoke often lacks the physical characteristics of real smoke, such as light scattering, transparency, and diffusion irregularities, and cannot simulate the interaction between the complex background environment and smoke, which can lead to a questionable generalization ability of the model if it overly relies on synthetic data for training (Park et al., 2022), and an increased risk of the evaluation metrics being invalidated in the real data. The samples in a dataset (Wu et al., 2023) often determine the patterns, features, and laws that a model can learn. In order to allow the model to effectively learn and adapt to a variety of different real-world forest scenarios, we carefully constructed a collection of 3352 high-quality visual visible-light images collected through field photography. This collection includes both smoke samples and false alarm smoke samples. In addition to this, we paid special attention to collecting images under different conditions, covering multiple viewpoints, different weather conditions (e.g., sunny, rainy, snowy), and variations in lighting conditions (from early morning to night). Images were taken in the field, from planned burn sites in Kunming, Yunnan Province, China, real-time monitoring in rural areas of Sichuan and Chongqing, and burn scenes from fire training around Beijing. The fires at these sites were tightly controlled, with burning operations conducted under controlled conditions, ensuring that smoke samples were collected in a safe environment that was as close as possible to real fire scenarios. Some of the samples and label distributions of the visual image set are shown in Fig. 1.

3.1.2. Fire risk factor data

Major forest fires often occur under three basic conditions: a large amount of dry, continuously distributed combustible material, an ignition source with a certain energy intensity, a driving force for the spread of the fire, and an adequate oxygen supply (Jones et al., 2022). For example, the Mediterranean region (Prema et al., 2022) is prone to forest fires in the summer because it often meets the conditions of the "Megafire Triangle" or the "30-30-30 Rule", i.e., surface temperatures exceeding 30 °C, relative humidity below 30 %, and winds exceeding 30 km per hour. Previous research (de Venancio et al., 2023) has identified a number of generalizable factors that are applicable to forest fire susceptibility modeling and mapping and are not restricted to specific regions. Ten of these factors were ranked among the top five predictors due to correlations higher than 0.94, including key variables such as temperature, wind speed, and precipitation. Our study specifically incorporated four measurements that are closely related to fire risk: temperature, wind speed, humidity, and 12-h precipitation, providing additional credibility to the use of imagery for monitoring and warning. In the construction of the risk assessment model, certain attributes were excluded from the dataset due to their inherent characteristics: firstly, sparse features, such as historical fire records, which are generally absent in non-fire areas, leading to insufficient reliability as predictive factors; secondly, region-dependent features, such as steep terrain, which can accelerate fire spread but whose influence significantly varies with the geographical environment. In flat areas, its effect is considerably weaker compared to mountainous regions. To avoid the potential limitation of these factors on the model's global generalization ability, this study selected meteorological factors that are strongly correlated, readily available in real-time, and universally applicable, as the core feature set, to balance the model's prediction accuracy and practical value.

3.1.3. Data-driven methodology

Fig. 2 covers the whole process from data collection to model training and evaluation. Specifically, the construction of the dataset consists of two main types of data sources: one is smoke visual data, which was collected through multi-angle images covering oblique, horizontal, and aerial views, providing diverse inputs to the model from different spatial dimensions. The second is fire risk factors that are closely related to fire risk, such as temperature (T), wind speed (WS), humidity (H), precipitation (Ppt), and smoke information (Smk), which are dynamically captured by on-site sensors and historical records,

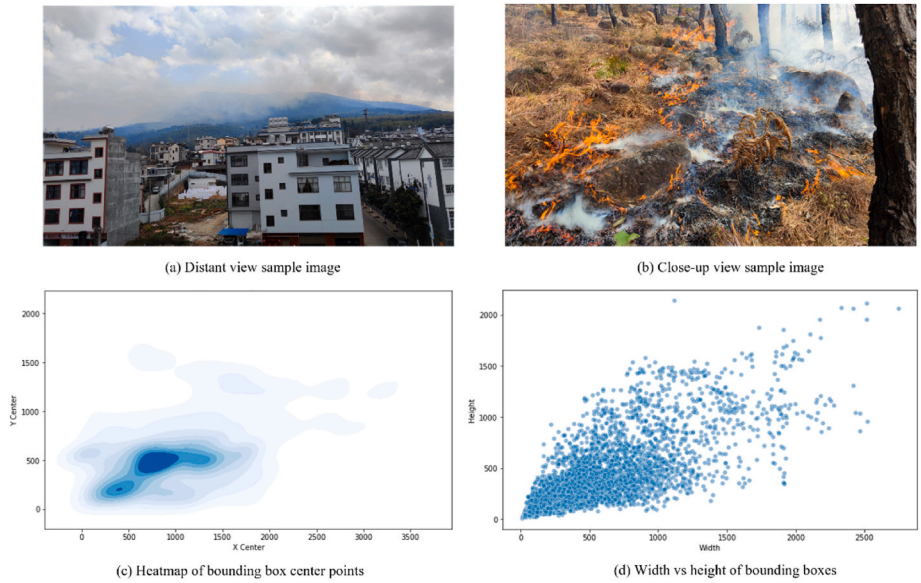

Fig. 1. Sample images from the dataset and the label distribution.

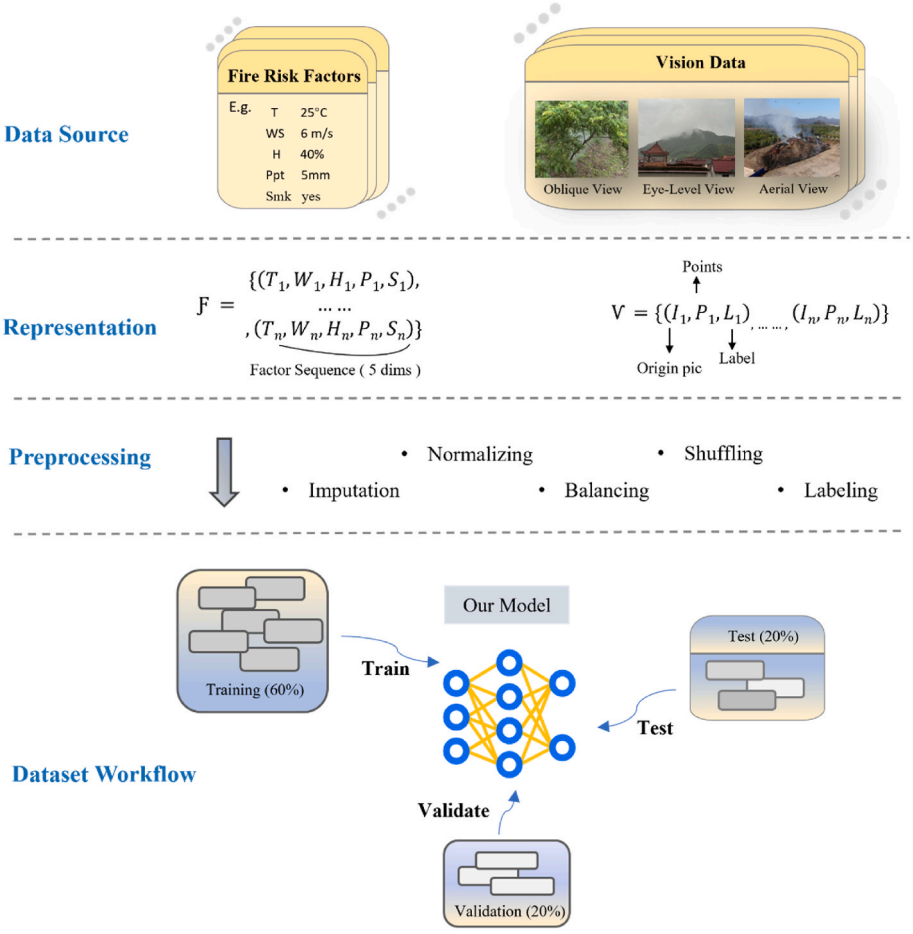

Fig. 2. Diagram of integrated dataset pipeline.

providing rich environmental characteristics.

The visual data part includes raw images, corresponding categories (Label), and labeled bounding boxes (Bbox) for accurately categorizing and locating smoke, while the fire hazard data is represented as a five-dimensional numerical feature vector containing key variables such as temperature, humidity, wind speed, precipitation, and smoke information, where smoke information is automatically generated by the target detection side. The two were spatially and temporally linked to generate a multimodal database.

The method for constructing fire risk ground truth involves determining sample risk levels through a dual mechanism of scenario patterns and environmental state constraints. High-risk samples (accounting for 53.2 %) are determined based on two strictly controlled scenarios: (1) Real fire events confirmed by firefighting experts, with meteorological station data collected from within a 5-km radius of the fire site through emergency management departments; (2) Planned burn experiments conducted in collaboration with local forest and grassland authorities and firefighting units, where fire intensity is controlled to ensure complete combustion and exclude out-of-control risks (fire spread speed <1 m/min, combustion efficiency >80 %). Low-risk samples (accounting for 46.8 %) are constructed based on two counterfactual scenarios: (1) Non-fire-origin smoke simulations, including water vapor, industrial emissions, and other false positive scenarios; (2) Smoke observations under extreme weather conditions, such as after rain, snowfall, or fog. Even though visual smoke characteristics may be present, the overall meteorological factors (temperature <10 °C, humidity >80 %) suppress the risk to a low level. For labeling, three forest fire prevention experts and two Ph.D. students were invited to classify each sample based on the "Forest Fire Weather Risk Level" (GB/T 36,743–2018), without knowing the sample sources. Finally, Krippendorff's alpha (α) (Krippendorff, n.d.) coefficient was used to test the labeling consistency, and with $\alpha = 0.892$, the consistency is higher than the acceptable threshold.

We have denoised, enhanced, and class-balanced the image data and made data label pairing. Given the limited scale of originally collected samples, we designed a combinatorial augmentation strategy for raw images, as detailed in Table 1. The geometric augmentation applies physically constrained spatial transformations to preserve fire dynamics

Table 1
Enhancement strategies for image data.

Augmentation Tier	Combination Strategy	Generated Qty	Application Notes
Geometric Transformations	Random Horizontal Flip + Restricted Rotation Angles	1205	Basic spatial diversity enhancement
	Random Horizontal Flip + Random Cropping	987	Prevent key feature loss
Photometric Adjustments	Brightness Adjustment + Gaussian Noise Injection	1632	Simulate illumination& signal decay
	Brightness Adjustment + Motion Blur	1074	Dynamic motion blur Scenarios
Cross-Tier Combinations	Geometric + Photometric Full Combination	2475	Maximum complexity augmentation
Original Data	–	3352	Non-augmented baseline

plausibility, while the photometric augmentation simulates real-world surveillance photometric variations (e.g., illumination changes, sensor noise). Min-Max Normalization (Aksoy and Haralick, n.d.) was applied to fire risk numerical features to reduce the influence of different feature scales, thus improving the model's effectiveness in learning relationships between features. For missing values, we used linear interpolation to bridge the temporal resolution differences, and at the same time, we strictly screened and removed abnormal data in the preprocessing stage to ensure the integrity and consistency of the data.

Finally, the amount of preprocessed data was expanded to 10,698 pairs, which were randomly shuffled and disrupted and divided into training, validation, and test sets in the ratio of 6:2:2. The training set is used for the initial training of the model, the validation set is used to tune the hyperparameters and prevent overfitting, and the test set is used for final performance evaluation.

3.2. Overall architecture of MM-SRENet

In this work, we aim to detect smoke through images and recognize fire risk by fusing smoke scene images with fire risk data. Our proposed MM-SRENet framework, as shown in Fig. 3, consists of two parts: the target detection side and the recognition side.

The target detection end is responsible for smoke detection and adopts YOLOv8 as the base model, on which several improvements are made: We chose YOLOv8n, emphasizing that this version was selected for its smaller size to facilitate deployment on edge devices, ensuring efficient performance even on resource-constrained devices. Firstly, we introduce a Multi-Scale Dilation Block (MSDBlock) with stronger multi-scale sensing and detail capturing capabilities in the C2f block. Second, in the detection head part, a new detection head is added based on the structure of the original detection head to enhance the model's adaptive detection capability for abnormal morphological smoke. These two improvements enable the model to adapt more flexibly to the changing characteristics of smoke in different scenes. In addition, Repulsion Loss is introduced in the loss function to encourage separation between different targets, especially in complex situations such as smoke obscured by trees. The lightweight and efficient Starnet is used at the recognition end to perform feature extraction on the smoke scene images, and the extracted features are fused with the fire risk data and smoke information obtained at the detection end at a later stage, which enables it to effectively capture the relationships and dependencies between various data modalities.

The MM-SRENet architecture is shown in Fig. 3 The target detection end consists of three main parts: backbone network, neck, and head. The backbone network uses an advanced convolutional structure to perform basic feature extraction on the input image and generate feature maps at different scales layer by layer. The neck enhances the complementarity of features at different levels by fusing the semantic information of high-level features with the spatial details of low-level features through up-sampling, splicing, and C2f modules. In addition, the architecture contains 4 detection heads, a new one compared to the conventional design, which provides the model with more granular options to focus more on targets in a specific size range. The smoke information obtained at the detection end is one-dimensional and, together with our fire risk factor, constitutes another modal data in addition to the image, which is subsequently fed into the recognition network. In the specific image processing of the smoke scene, the recognition network first extracts the abstract features of the image through a backbone network based on star operation (Ma et al., n.d.). Next, in the feature fusion process, we use progressive encoder designfeature expansion and splicing operations to connect the information of different modalities at a specific level. This feature fusion strategy integrates the image data with the data streams of other modalities, allowing the model to capture the synergies and interdependencies between different modalities, and thus construct a more comprehensive understanding of multimodal inputs. Finally, the nonlinear fitting is implemented through a multilayer perceptron (MLP) to generate the output, i.e., the classification of fire risk modes. In this way, the final fire hazard recognition information becomes truly reliable and no longer relies solely on the presence of smoke to be judged as dangerous.

3.2.1. Multi-Scale Dilation Block

We propose to improve the C2f module in the backbone network at the smoke detection end, and Fig. 4 shows the structure of C2f-MD and its details. Specifically, we add our own Multi-Scale Dilation Block, a multi-scale feature extraction module, after its last convolutional layer. The core of the MSDBlock module is the dilation rate stratification strategy (dilation_rate = [1, 3, 5]), which uses multiple parallel convolution branches to extract features at different scales and receptive fields. This approach leverages residual connections to enhance gradient propagation, thereby improving the model's ability to perceive complex targets such as smoke. MSDBlock is designed to capture sufficient receptive fields without altering the image resolution, allowing the network to focus on multi-scale contextual semantic information. This is particularly important for smoke detection, as smoke exhibits multi-scale characteristics (such as small-scale blurry edges, medium-scale texture variations, and large-scale region coverage). By adjusting the dilation rates, the receptive field can be flexibly expanded to adapt to the varying scales of smoke features, thereby improving the model's performance in complex forest environments. It first uses a 1×1 convolution (Conv1x1_1) to scale the input to the intermediate number of channels, mid_channels, and performs initial feature fusion with batch

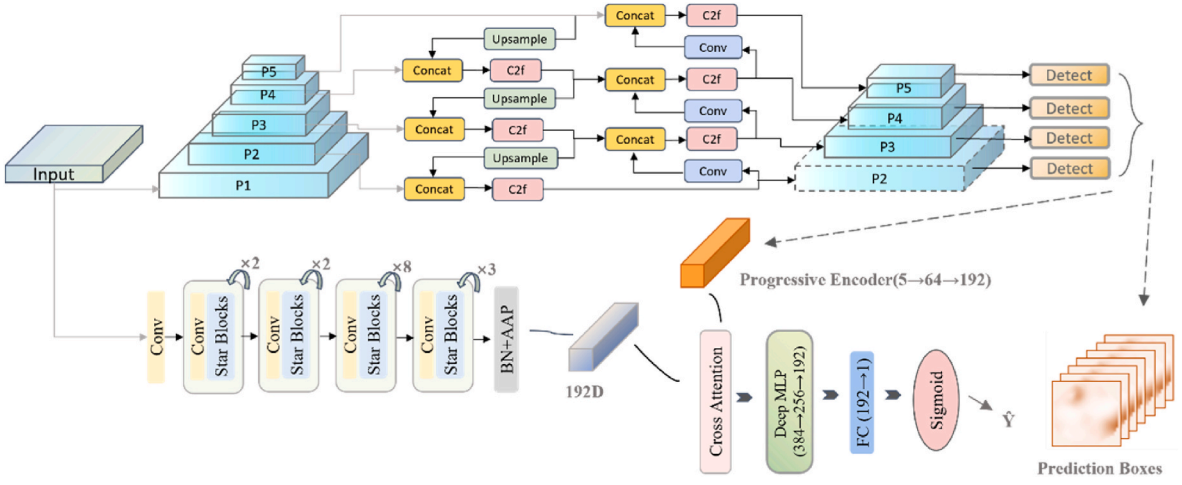

Fig. 3. The overall architecture of MM-SRENet.

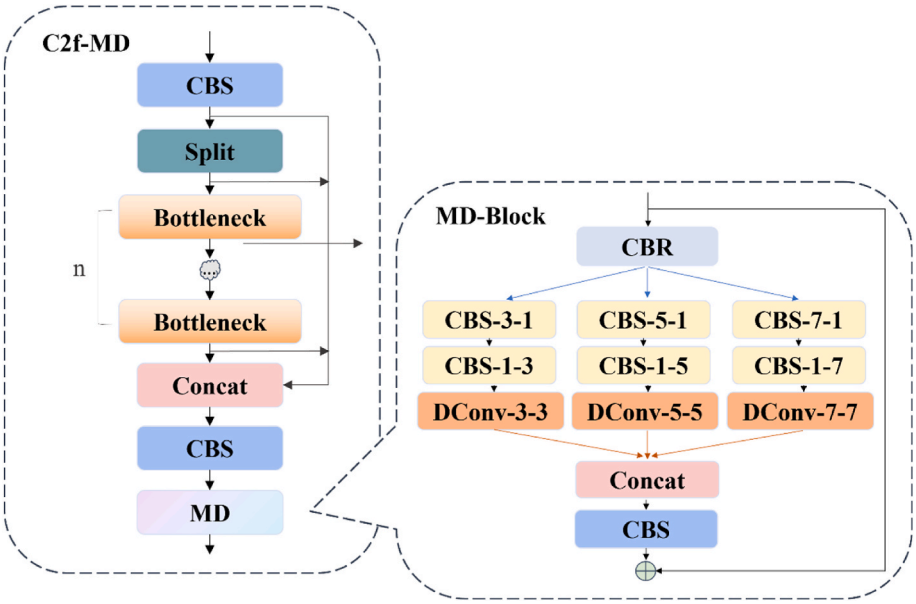

Fig. 4. Structure of C2f-MD.

normalization and ReLU activation functions. The features are then fed into three parallel convolutional branches, each with a different convolutional kernel size and nulling rate: the first branch (conv1) uses (3,1), (1,3) convolution and (3,3) nulling convolution for extracting small-scale features; the second branch (conv2) uses (5,1), (1,5) convolution and (5,5) nulling convolution for extracting intermediate-scale features The third branch (conv3) uses (7,1), (1,7) convolution and (7,7) null convolution to extract large scale features. The feature maps output from the three branches are spliced in the channel dimension to form a feature map containing multi-scale information, and then another 1×1 convolution (Conv1x1_2) is used to achieve feature fusion and computational control, and the batch normalization and SiLU activation functions are used to further stabilize the fused features and to enhance the nonlinear representation. Finally, the module employs residual concatenation to sum the fused features with the original input, which preserves the input information and enhances the feature representation.

3.2.2. Optimization of shallow feature

Forest smoke detection tasks usually involve more complex natural environments, e.g., in bright daytime, with low contrast between the background and the target area, irregular shape, and non-uniform density of the smoke, when the detection needs to rely on the accurate capture of small and blurred features in the image. Since smoke usually shows slight gray or white atomized features in small areas, we add a shallow feature map P2 layer detection head on original three-layer feature maps (P3, P4, and P5), as shown in Fig. 3. The P2 layer has a higher resolution and is more sensitive to the edges of the smoke, subtle textures, and other details, which effectively improves the ability to capture and perceive the small-scale features of the smoke. The P2 layer has higher resolution and is more sensitive to details such as edges and fine textures, which can effectively improve the ability to capture and perceive small-scale smoke features. In Neck, we sequentially fuse the P2 features with the P3, P4, and P5 feature maps through up-sampling and concat operations. This structural cascade ensures the gradual propagation of shallow features through all levels of feature maps, and realizes the fusion of details and high semantic information.

3.2.3. Selective Repulsion Loss

The original loss function consists of Classification Loss, Bounding Box Regression Loss, and Distribution Focal Loss to balance the classification and localization needs in the detection task.

$$\mathscr{L} = \lambda_{box} \cdot \mathscr{L}_{box} + \lambda_{box} \cdot \mathscr{L}_{cls} + \lambda_{dfl} \cdot \mathscr{L}_{dfl} \tag{1}$$

The default weights of the loss function are:

$$\lambda_{box} = 7.5 \tag{2}$$

$$\lambda_{cls} = 0.5 \tag{3}$$

$$\lambda_{dfl} = 1.5 \tag{4}$$

However, in a single-category smoke detection task, DFL is less useful and can be ignored for the time being. Smoke detection has special needs, the targets usually have fuzzy boundaries and high uncertainty, the focus of this task is to accurately locate the boundaries of the smoke region, and there is no need to consider the category imbalance among multiple categories. At the same time, we introduce Repulsion Loss (Wang et al., 2017) as a new optimization term to enhance the bounding box discrimination between adjacent targets. By applying a "repulsion force", the repulsion loss avoids excessive overlapping of bounding boxes when detecting neighboring smoke regions, and thus reduces the interference of the prediction results. Specifically, the repulsion loss allows the model to focus more on distinguishing between neighboring smoke regions when generating the bounding box, thus reducing the occurrence of false detections and boundary blurring. This design is particularly suitable for scenarios where smoke is dispersed in complex backgrounds, such as fire monitoring where different smoke clumps are close to each other or even partially overlap when smoke gradually spreads, or smoke spreads in a forest and forms highly confusing images in an environment interspersed with foliage and shadows.

Our design is as follows:

$$\mathscr{L} = \lambda_{box} \cdot \mathscr{L}_{box} + \lambda_{cls} \cdot \mathscr{L}_{cls} + \lambda_{repGT} \cdot \mathscr{L}_{repGT} + \lambda_{repBox} \cdot \mathscr{L}_{repBox} \tag{5}$$

In addition to the initial weights mentioned earlier, the newly introduced λ_{repGT} and λ_{repBox} act as balancing weights for the auxiliary losses, with values of 0.5 and 2.0, respectively.

The purpose of the repGT Loss is to prevent a candidate box from overlapping too much with non-target ground truth boxes. Specifically, for a given candidate box P, the objective is to minimize the intersection ratio (IoG) between P and the nearest non-target ground truth box. This is achieved by penalizing the overlap between the candidate box B^P and

the excluded ground truth box G^P_{Rep}. The *SmoothIn* function, a smooth logarithmic function, adjusts the sensitivity of the rejection loss to outliers. As the candidate box overlaps more with the non-target ground truth box, the repGT loss value increases, effectively suppressing the candidate box from drifting towards the non-target ground truth box.

$$\mathscr{L}_{repGT} = \frac{1}{|P_+|}\sum_{P\in P_+}\text{Smooth}_{ln}\left(IoG\left(B^P, G^P_{Rep}\right)\right) \tag{6}$$

The goal of repBox Loss is to reduce the overlap between candidate boxes of different objects before the Non-Maximum Suppression (NMS) step, thereby lowering the risk of incorrect merging of candidate boxes. First, the candidate box set P_+ is divided into non-overlapping subsets by object category. For two candidate boxes B^P_i and B^P_j of different categories, their overlap should be as small as possible. A small constant ϵ is introduced to ensure that the denominator does not become zero. This loss term reduces the overlap between candidate boxes of different categories, making the detector more robust during the NMS process and decreasing the likelihood of incorrect merging.

$$\mathscr{L}_{repBox} = \frac{\sum_{i\neq j} Smooth_{ln}\left(IoU\left(B^P_i, B^P_j\right)\right)}{\sum_{i\neq j} 1\left[IoU\left(B^P_i, B^P_j\right) > 0\right] + \epsilon} \tag{7}$$

In this way, the design of Selective Repulsion Loss provides a more targeted optimization approach for our task, playing an effective role in measurement, guidance, and adjustment during model training.

3.2.4. Star block

In recognition tasks, extracting representative smoke scene patterns from the original image is the critical first step. To achieve this, we utilized the Star Block structure in the backbone network, with its internal architecture illustrated in Fig. 5. The core innovation of the Star Block lies in the element-wise multiplication (*) operation, a specialized matrix computation method capable of implicitly mapping low-dimensional input features to high-dimensional nonlinear feature spaces. Considering the complexity and nonlinearity of smoke in terms of shape, texture, and distribution, this operation is not only more efficient but also exhibits superior representational capacity compared to traditional feature fusion methods.

The internal structure of the Star Block primarily consists of depthwise convolution layers (DW), fully connected layers (FC), and nonlinear activation functions (such as ReLU6). First, the input features are processed by the depthwise convolution layer (DW) to capture the texture information of smoke in different regions. Then, two fully connected layers (FC1 and FC2) are employed to expand and compress the channel-wise features, simulating complex channel interactions and extracting deeper semantic features. The element-wise multiplication operation is used to fuse the two-channel feature branches, avoiding information loss while further enhancing the representation of higher-order features.

Moreover, to improve the network's stability and generalization ability, the Star Block is designed with residual connections and DropPath regularization mechanisms. Residual connections effectively retain the original input information and optimize gradient flow in deeper networks, while DropPath enhances the model's robustness to background noise in complex scenarios. This organic combination achieves hierarchical feature extraction across spatial and channel dimensions, balancing efficiency and accuracy.

3.2.5. Multimodal fusion recognition

Based on the Starnet backbone network, hierarchical feature extraction is performed through four stages (Stages), each containing 2, 2, 8, and 3 Star Block modules, respectively. After global average pooling, a 192-dimensional image feature vector $F_{img} \in \mathbb{R}^{192}$ is generated as the first step of recognition.

The presence of smoke is also obtained from the detection end. Combined with additional fire risk data $X_{img} \in \mathbb{R}^5$, the model further identifies fire risk. First, the risk data is projected to 64 dimensions via a fully connected (FC) layer with ReLU activation, then mapped to a 192-dimensional feature vector $F_{risk} \in \mathbb{R}^{192}$ through a secondary alignment step. This progressive encoder design mitigates overfitting risks.

A dual-stream interaction mechanism is introduced, where L2 normalization is first applied to F_{img} and F_{risk}, respectively, followed by the computation of cross-modal dependency weights via the cross-attention module:

$$Q = W_q F_{\text{img}} \tag{8}$$

$$K = W_k F_{\text{risk}} \tag{9}$$

$$V = W_v F_{\text{risk}} \tag{10}$$

$$\text{Attn}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{11}$$

Here, $W_q, W_k, W_v \in \mathbb{R}^{192\times192}$, are learnable parameters, and $d = 192$ is the scaling factor. The attention output is combined with the original features via residual connection and fed into a deep MLP (structure: 384 → 256→192, with BatchNorm-ReLU-Dropout) to obtain the fused feature $F_{fusion} \in \mathbb{R}^{192}$.

Finally, a binary classification head outputs the risk probability:

$$P(y=1|X) = \sigma\left(W_c \cdot \text{LayerNorm}\left(F_{\text{fusion}}\right) + b_c\right) \tag{12}$$

Where σ is the Sigmoid function, $W_c \in \mathbb{R}^{1\times192}$ is the classification weight, and LayerNorm stabilizes training dynamics. The loss function employs label-smoothed binary cross-entropy (smoothing factor = 0.1) with gradient clipping (max_norm = 1.0).

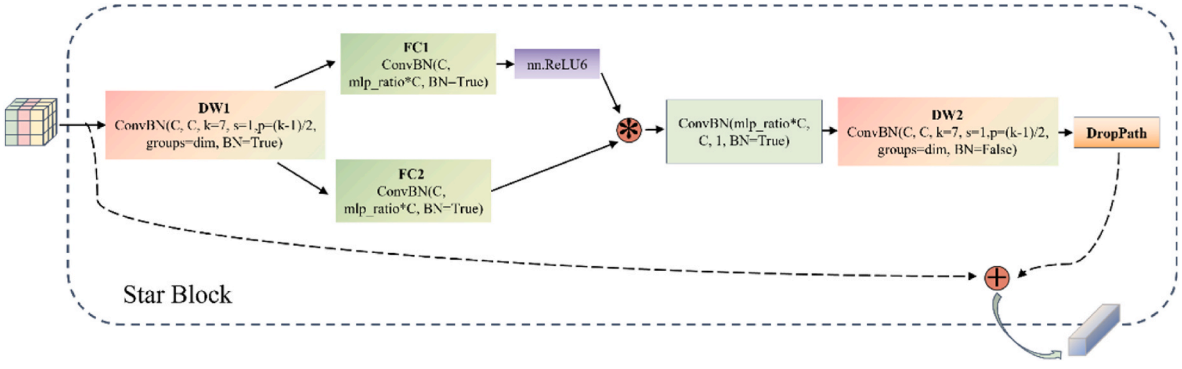

Fig. 5. Structure of star block.

4. Experimentation and analysis

4.1. Experimental setup

The training and testing were conducted on a dedicated server with an Intel Xeon W-2223 CPU (8 logical cores) and two GPUs: Quadro RTX 5000 (16 GB VRAM) and Quadro RTX 4000 (8 GB VRAM). The models were implemented using PyTorch 1.13.0 and Python 3.8.20. All comparative and ablation experiments were conducted on the same platform to ensure fairness and consistency.

4.2. Evaluation indicators

This study employs multi-dimensional evaluation metrics for a comprehensive assessment of model performance. In the smoke detection task, precision (P), recall (R), mean average precision (mAP, IoU = 0.5), and F1-score are used to measure the model's overall prediction accuracy, false positive control ability, false negative rate, and overall performance under class imbalance. The above metrics can be calculated using the following formulas:

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$AP = \int_0^1 P(R)d(R) \tag{15}$$

$$F_1 = \frac{2*P*R}{P + R} \tag{16}$$

TP represents the number of samples correctly predicted as positive, FP represents the number of samples incorrectly predicted as positive, FN represents the number of samples that are actually positive but incorrectly predicted as negative, and TN represents the number of samples correctly predicted as negative.

In the fire risk assessment task, Accuracy (ACC) is used to measure the overall correctness of the model, and the Risk Score Error (RSE) quantifies the deviation between the predicted fire risk score and the actual risk. These metrics can be calculated using the following formulas:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$RSE = \frac{1}{N}\sum_{i=1}^{N} |P_i - y_i| \tag{18}$$

Here, P_i represents the predicted risk score for the i-th sample, y_i represents the true risk score for that sample, and N is the total number of samples. A lower RSE value indicates that the model's risk score predictions are more accurate.

Given the real-time responsiveness requirements of forest fire early warning and the limitations of the monitoring system's hardware resources, we use detection delay (seconds) and risk assessment delay (seconds) to measure the response time of smoke detection and risk scoring, respectively. We prioritize model designs with lower computational complexity (GFLOPs) and fewer parameters.

In addition, the robustness of the model is verified through False Positive Rate (FPR) and Environmental Adaptability (EA). The False Positive Rate is used to assess the model's false alarm level in non-fire scenarios, while Environmental Adaptability quantifies the model's performance fluctuations across different environments. The formulas we define are as follows:

$$FPR = \frac{FP}{FP + TN} \tag{19}$$

$$EA = 1 - \frac{\text{Performance}_{max} - \text{Performance}_{min}}{\text{Performance}_{max}} \tag{20}$$

Performance_{min} and Performance_{max} represent the model's performance in the worst and best environments, respectively.

The above metrics comprehensively evaluate the model's practical application performance in terms of accuracy, real-time responsiveness, and robustness.

4.3. Performance of the smoke detection

4.3.1. Training strategy

The model training used the AdamW optimizer, with a training epoch of 100 and a batch size of 16. The initial learning rate was set to 0.01, and a Cosine Annealing Scheduler was employed to dynamically adjust the learning rate, helping to stabilize convergence in the later stages of training. To reduce the false positive rate, pure background images were incorporated during training. Additionally, strong data augmentation was gradually reduced in the later stages of training to further improve detection performance.

4.3.2. Detection results

In Fig. 6, group (a) shows examples of the labeled dataset, representing the manually annotated true smoke regions in the validation set; group (b) displays the model's predictions for the corresponding samples. By comparing groups (a) and (b), the model's performance in locating and identifying smoke is clearly evident. In most examples, the model successfully detected the smoke regions marked by the labels, showing high robustness across different concentrations and forms. Additionally, the confidence scores shown in group (b) reflect the model's confidence in its predictions for each region. Even in more ambiguous areas, the model assigns some confidence, indicating its ability to distinguish subtle differences. Furthermore, the model also detected smoke in some regions that were not labeled, demonstrating its high sensitivity to smoke features that may have been overlooked by human annotators, as well as its generalization ability across different scenarios.

4.3.3. Ablation experiments

To validate the effectiveness of the proposed method in the smoke detection task, we designed a series of ablation experiments to gradually optimize and compare different modules of the model. Specifically, these include the added detection heads (Heads), C2fMD and its introduction at different positions, and the improved loss functions (Loss), among others. Based on these, we compare various performance metrics (P%, R%, mAP%, F1Score%) from different experimental results, which are detailed in Table 2.

The baseline model demonstrates limited effectiveness for small-scale blurry targets in complex forest environments, achieving 73.1 % precision and 75.5 % mAP. The introduction of specialized detection heads in Algorithm 2 yields a substantial 4.5 % precision improvement to 77.6 %, with comparative analysis against Algorithm 10 and Algorithm 11 confirming the heads' critical role in feature refinement. Systematic evaluation of C2f-MD module placement reveals Backbone integration as optimal, delivering 78.6 % mAP compared to 77.2 % and 77.8 % in Neck-only and Full-network configurations respectively. The combined application of detection heads and C2f-MD modules in Algorithm 7 elevates performance to 80.3 % precision and 79.5 % mAP, establishing Backbone optimization as the most cost-effective strategy. Repulsion Loss implementation demonstrates measurable boundary separation improvements, particularly enhancing smoke differentiation as evidenced by Algorithm 13's 78.5 % mAP. Our final configuration in

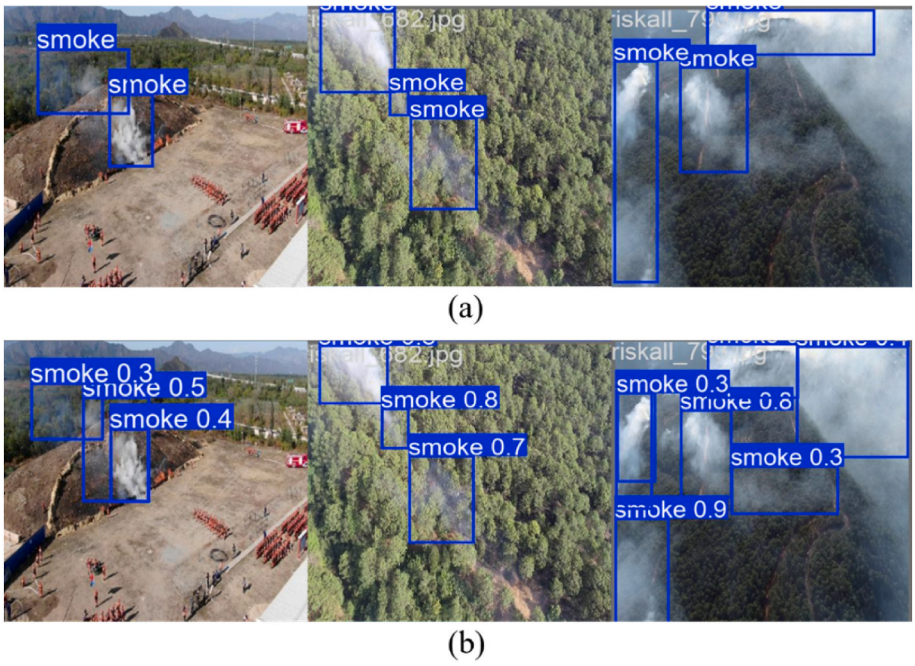

Fig. 6. Visualization of labels and real detectiom results: (a)Original labels; (b)real detectiom results.

Table 2
Results of ablation experiments.

Algorithm	Heads	C2f-MD			Loss	Evaluation indicators			
		Backbone Only	Neck Only	Full		P%	R%	mAP%	F1Score%
1 (baseline)						73.1	67.2	75.5	70.03
2	√					77.6	69.7	79.1	73.44
3		√				78.2	69.4	78.6	73.54
4			√			78.1	68.9	77.2	73.21
5				√		78.5	69.8	77.8	73.89
6					√	78.2	67.5	76.9	72.46
7	√	√				80.3	73.1	79.5	76.53
8	√		√			79.7	72.5	78.8	75.93
9	√			√		80.5	72.3	79.3	76.18
10	√				√	77.0	69.6	77.9	73.11
11		√			√	79.3	70.5	78.1	74.64
12			√		√	78.9	69.3	77.4	73.7
13				√	√	79.0	69.6	78.5	74.00
14	√		√		√	79.4	69.5	79.9	74.12
15	√			√	√	79.6	71.7	80.3	75.44
16(Ours)	√	√			√	**81.6**	**72.9**	**81.2**	**77.01**

Algorithm 16 synergistically integrates all optimizations, achieving state-of-the-art performance with 81.6 % precision and 81.2 % mAP. The 5.7 % mAP gap between Algorithm 16 and the baseline quantitatively validates the cumulative effectiveness of our architectural improvements.

To validate the efficacy of our Selective Repulsion Loss design, we systematically analyze the impact of hyperparameters λ_{repGT} and λ_{repBox} on smoke detection performance, as detailed in Table 3. The baseline model, Algorithm 1 with λ_{repGT} and λ_{repBox} both set to 0, achieved an mAP of 79.5 % and an F1 score of 76.53 %. When introducing λ_{repBox} with an equal weight of 0.5 in Algorithm 2, the mAP decreased by 1.1 %, indicating the limitations of uniform weighting for this task. By gradually increasing the λ_{repBox} weight, as shown in Algorithms 3 to 5, a clear trend emerged: higher λ_{repBox} weights enhanced geometric consistency in bounding box regression, improving mAP from 79.6 % in Algorithm 3–81.2 % in Algorithm 5, while maintaining a precision of 81.6 % and an F1 score of 77.01 %. However, overemphasizing λ_{repGT}, as in Algorithm 6 with λ_{repGT} set to 1.0, caused the mAP to drop to 80.1 %, highlighting the need for dynamic balance between parameters. The optimal configuration, Algorithm 5 with $\lambda_{repGT} = 0.5$ and $\lambda_{repBox} = 2.0$, validates the core value of task-oriented weight allocation. By reducing reliance on Non-

Table 3
The ablation study on λ_{repGT} and λ_{repBox}

Algorithm	λ_{repGT}	λ_{repBox}	P%	R%	mAP%	F1Score%
1 (baseline)	0.0	0.0	80.3	73.1	79.5	76.53
2	0.5	0.5	79.9	73.4	78.4	76.51
3	0.5	1.0	80.7	72.7	79.6	76.49
4	0.5	1.5	81.3	73.6	80.5	73.96
5(Ours)	0.5	2.0	**81.6**	**72.9**	**81.2**	**77.01**
6	1.0	2.0	79.5	72.3	80.1	75.73
7	0.3	1.5	79.8	72.5	79.7	75.98

Maximum Suppression post-processing and enhancing localization refinement, this configuration significantly improves smoke detection performance, particularly in scenarios sensitive to minor spatial deviations under IoU-based evaluation. These results empirically support the necessity of the asymmetric weighting strategy, providing theoretical grounding for addressing both post-processing sensitivity and class imbalance challenges.

Additionally, although our optimal combination achieved a more than 5 percentage point improvement over the baseline model, the mAP value still did not reach an exceptionally high level, which is closely related to the characteristics of the dataset. Due to the unclear boundaries of smoke regions and their varying shapes, the annotation process carries a significant degree of subjectivity. Particularly in large, irregularly shaped smoke areas, they may be divided into multiple annotated boxes, as shown in the label group of Fig. 6. This subjective segmentation leads to substantial variation in annotations, either between different annotators or by the same annotator at different times, reducing annotation consistency. Even though the model successfully detects the actual smoke locations, slight discrepancies between the detection results and annotations can lead to a lower score. In this case, despite the model's detection results being sufficiently accurate or even ideal in practical applications, the mAP value, constrained by strict validation rules, does not fully reflect the model's true detection capabilities.

4.3.4. Comparison with classical networks

The design philosophy of MM-SRENet emphasizes lightweight deployment and multimodal adaptability, achieving a well-balanced trade-off between computational efficiency and environmental perception. The detection module demonstrates significant advantages among models with the same parameter scale, as shown in Table 4: leveraging MSDBlock-enhanced hierarchical feature extraction and dual-optimized small-object detection heads, it effectively mitigates the challenges posed by the multi-scale nature of smoke targets. Notably, in typical forest fire monitoring scenarios, CNN-based detectors outperform Transformer architectures (e.g., ViT/DETR) in capturing small targets. Early smoke captured by watchtowers and UAVs commonly exhibits low contrast (average grayscale difference <15) and small spatial coverage (pixel ratio <5 %), where the global feature modeling advantage of Transformer-based models becomes a limitation. Consequently, their mAP50 performance drops by 4.4–5.1 percentage points in this task.

The design of the multimodal fusion module further validates the effectiveness of the proposed architecture. Experimental results in Table 5 show that our approach achieves breakthroughs in fire risk identification accuracy (93.7 % vs. 90.6 % in the next best approach) and false positive rate (7.1 % vs. 10.6 %), while maintaining a low inference latency of 0.028s. Comparative analysis reveals that while CNN + LSTM leverages temporal modeling for decent numerical performance (90.6 % ACC), its cascaded structure introduces 21 % computational redundancy. FCN + LSTM suffers from the lack of local feature modeling, resulting in a drop in accuracy to 85.1 %. Meanwhile, Transformer + MLP, despite its theoretical global perception capability, exhibits excessive parameter sensitivity (requiring large-scale datasets for convergence), making it less suitable for real-world deployment.

Table 4
Performance comparison of smoke detection models.

Methods	Backbone	Evaluation Metrics		
		Pixels	mAP50	Params(M)
Faster-RCNN	ResNet50	512	69.4	28
MobileNet-SSD	MobileNet	512	72.7	5.8
YOLOv5-n	CSPDarkNet	512	73.6	8.5
YOLOv7-n	ELANNet	512	78.9	6.2
DETR	ResNet50	512	76.5	42
ViT-Ti	–	512	77.2	5.7
Ours	Proposed	512	**81.6**	**7.4**

Table 5
Multimodal fusion approaches for fire risk assessment.

Methods	Modalities	Evaluation Metrics		
		ACC	FPR	Lantency(s)
CNN + LSTM	RGB + Sensor	90.4	10.6	0.034 +- 0.0012
FCN + LSTM	RGB + Sensor	85.1	14.8	0.054 +- 0.0008
ViP + LSTM	RGB + Sensor	86.4	15.5	0.041 +- 0.0052
ViP + MLP	RGB + Sensor	85.3	16.9	0.039 +- 0.0046
Transformer + MLP	RGB + Sensor	84.6	17.8	0.151 +- 0.0079
Ours	RGB + Sensor	**93.7**	**7.1**	**0.028 +- 0.0009**

This study proposes an innovative dual-stream interaction mechanism, where a CNN backbone designed with Star Operation extracts fine-grained spatial features, while a cross-modal cross-attention mechanism optimizes the fusion modeling of fire risk information. Finally, the fused features are passed through a deep MLP for nonlinear transformation. This "spatial-semantic" co-optimization strategy explicitly expresses the implicit correlations between smoke patterns and environmental parameters, ultimately achieving 93.7 % decision stability in complex forest fire risk assessment scenarios.

4.4. Results of the fire risk assessment

4.4.1. Correlation analysis

This study comprehensively assesses fire risk by integrating multimodal data. The correlation matrix heatmap in Fig. 7(a) clearly illustrates the relationships between five key variables: smoke area (Area), environmental temperature (Temp), wind speed (Wind), humidity (Humid), and precipitation (Precip). Among these, the smoke area is the output of the early-stage smoke detection network, while the remaining four variables are derived from environmental data for later-stage multimodal fusion recognition. Together, these variables form a comprehensive feature set for fire risk assessment.

The analysis reveals a significant negative correlation between environmental temperature and humidity (correlation coefficient = −0.62). This result confirms that high temperatures and low humidity are critical conditions for high fire risk, further indicating that the joint variation in temperature and humidity is a key focus in fire risk assessment. In addition, wind speed and humidity also exhibit a strong negative correlation (correlation coefficient = −0.64). Increased wind speed not only lowers humidity but also accelerates the spread and range of fires, significantly enhancing their destructive potential. Conversely, precipitation shows a positive correlation with humidity (correlation coefficient = 0.65), suggesting that precipitation plays a significant role in increasing air humidity and reducing fire risk. These findings underscore the importance of considering precipitation as a mitigating factor in fire risk.

As an important indicator of fire, the analysis of the correlation between smoke area and environmental variables reveals the interactions between different factors. Specifically, there is a mild negative correlation between smoke area and wind speed (correlation coefficient = −0.33), indicating that in high-wind environments, smoke disperses rapidly and struggles to form high-concentration regions. Additionally, the positive correlation between smoke area and precipitation (correlation coefficient = 0.32) suggests that precipitation may suppress the fire and cause the smoke to accumulate in localized areas.

The above analyses provide an important basis for the design of the network input. Additionally, the unified manifold approximation and projection (UMAP) in Fig. 7(b) illustrates the dimensionality reduction effect of high-dimensional data, with different colors representing different label clusters. Overall, the distinction between fire risk and non-fire risk data is quite clear. Even after mapping the high-dimensional features to a two-dimensional space, data with different fire risk statuses can still be effectively separated. Some categories

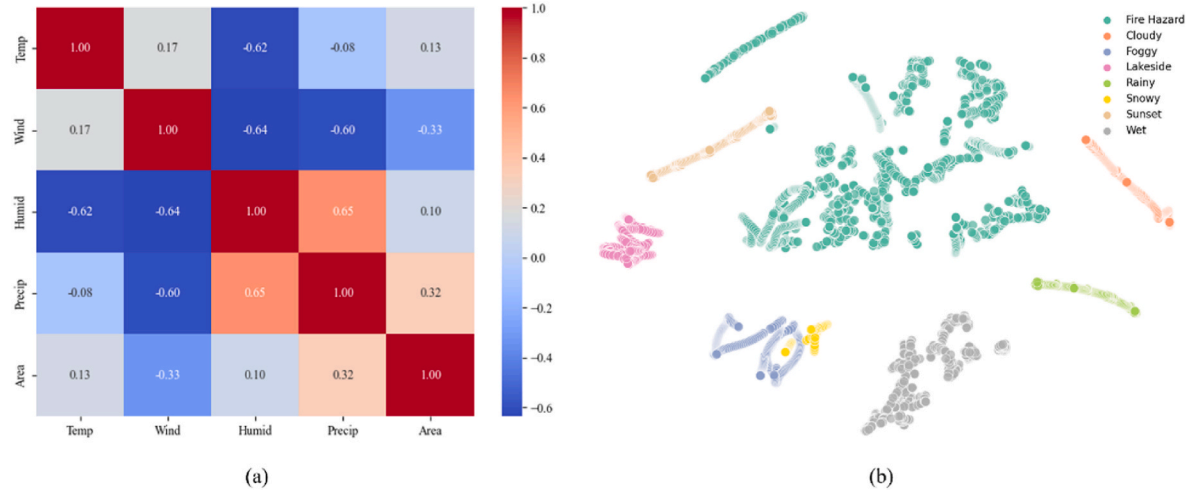

Fig. 7. Correlation matrix and UMAP visualization of multimodal fire risk data.

exhibit gradual changes in features, which are often the result of continuous weather changes and the interaction of multiple variables. It is noteworthy that the fire-risk data shows a more dispersed distribution, indicating the diversity of features involved in fire-risk situations. However, despite the dispersion, these data still form a distinct separation from non-fire risk data, providing a clear basis for fire risk monitoring and early warning.

The complex interrelationships among the multimodal data mentioned above suggest that the non-linear effects and interactions of environmental variables need to be carefully considered when building the network model. By incorporating these variables and their correlations into the fire risk assessment model, more scientifically grounded data support can be provided for fire early warning and prevention.

4.4.2. *Statistical robustness and feature importance analysis*

To investigate the contribution of fire risk factors to predictive outcomes, we conducted a comprehensive machine learning analysis utilizing the XGBoost framework coupled with SHAP (SHapley Additive exPlanations) interpretability methods. The dataset was initially partitioned into training and testing subsets at an 8:2 ratio to ensure unbiased performance evaluation. Hyperparameter optimization was systematically executed through grid search (GridSearchCV) with 5-fold cross-validation ($k = 5$), where the dataset was iteratively divided into five stratified subsets. For each iteration, four subsets were allocated for model training while the remaining subset served for validation, ensuring robust generalization assessment. Model selection prioritized minimization of logarithmic loss (log loss), a probabilistic accuracy metric that quantifies prediction uncertainty. The optimized hyperparameter configuration is detailed in Table 6, demonstrating significant improvements in predictive calibration. To statistically validate model reliability, we report both the mean and standard deviation of cross-validated performance metrics in Table 7. The constrained variability ($SD < 0.015$ across folds) confirms model robustness, indicating consistent wildfire risk prediction capability across heterogeneous data partitions.

Table 6
The optimized hyperparameters.

Parameters	Scope of search	Optimum value
n_estimators	[100, 200, 300, 400, 500]	500
max_depth	[3, 4, 5, 6, 7]	7
learning_rate	[0.01, 0.02, 0.05, 0.1]	0.05
colsample_bytree	0.6	0.6
subsample	0.7	0.7
booster	gbtree	gbtree
eval_metric	logloss	logloss

Table 7
Cross-validation results.

Cross-validation folds	Mean_test_score	Std_test_score
Fold 1	0.00284	–
Fold 2	0.00148	–
Fold 3	0.00381	–
Fold 4	0.01678	–
Fold 5	0.00224	–
Total	0.00543	0.004829

Furthermore, SHAP-based beeswarm visualization as Fig. 8 was employed to decode feature attribution patterns, where horizontal dispersion of data points represents the magnitude and directionality of each factor's impact on model outputs. Color gradients encode original feature values, revealing critical nonlinear relationships - for instance, the positive correlation between extreme temperature anomalies (red-clustered points) and elevated fire risk probabilities. The analysis reveals that temperature has the strongest influence on fire prediction. Similarly, humidity exhibits a strong correlation with the model's predictions, particularly as increased humidity raises the moisture content of combustible materials, thereby reducing their likelihood of ignition. This phenomenon aligns with the actual fire occurrence mechanism, as dry environments are more prone to fire outbreaks. Precipitation and wind speed have relatively smaller impacts but still play a regulatory role. Higher precipitation is generally associated with negative Shapley values, as stored moisture from rainfall can help reduce fire risk to some extent. The effect of wind speed is more complex—while higher wind speeds (red points) may, in some cases, increase fire risk by accelerating fire spread, their overall impact is not as significant as that of temperature and humidity. By incorporating SHAP analysis, we provide a transparent and interpretable understanding of the model's behavior, ensuring that the identified key factors align with domain knowledge in fire risk assessment.

4.4.3. *Comparative analysis with single-modal method*

To evaluate the performance of different source data in fire risk assessment, we selected evaluation metrics such as ACC, FPR, RSE, and EA for comparison, with detailed results presented in Table 8. These metrics have been thoroughly introduced earlier in the text.

The unimodal visual model relies solely on smoke images. While it can reflect fire scenarios to some extent, its lack of background prior knowledge results in an ACC of only 77.69 %, indicating weak risk discrimination ability. To address this limitation, the proposed

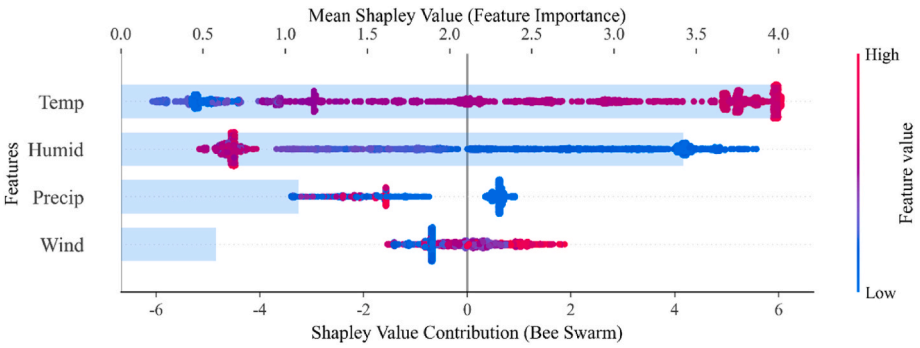

Fig. 8. Biaxial SHAP-based visualization combining beeswarm plot and feature importance plot.

Table 8
Model performance with various modalities.

Models	ACC%	FPR%	RSE	EA
Vision_only	77.69	22.97	0.31	0.82
Fusion (Final)	93.68	7.09 (-69.13 %)	0.15 (-51.6 %)	0.91

multimodal fusion model, MM-SRENet, incorporates key fire risk factors with smoke images. This strategy significantly improves the model's ACC to 93.68 %, greatly enhancing its ability to identify fire risks. Notably, the false positive rate (FPR) is closely related to the ACC metric. With only visual data, the FPR stands at 22.97 %, while introducing additional information reduces it dramatically to 7.09 %. Our fusion model can more accurately distinguish real fire scenarios from non-fire scenarios, effectively reducing false positives. For forest fire early warning systems, FPR is a critical metric.

The core of our task is to accurately assess the likelihood of fire occurrence rather than merely performing a simple binary classification. More importantly, precise risk scoring enables differentiation among various levels of fire risk. Given the diverse and dynamic nature of forest environments, RSE was introduced as an evaluation metric to analyze the error distribution under different spatiotemporal conditions. The fusion model achieves an RSE of 0.15, significantly outperforming the visual model's RSE of 0.31, reducing error by approximately 51.6 %. This indicates that the fusion model provides more accurate fire risk scoring.

Finally, in terms of the model's environmental adaptability, the EA metric shows an improvement of nearly 10 %, demonstrating greater robustness in handling the variability of forest environments. This adaptability is especially crucial for complex forest fire scenarios, ensuring that the model can consistently provide accurate risk assessments under varying external conditions.

4.4.4. *Interpretability analysis*

To enhance interpretability, we introduced Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., n.d.). This method visualizes the weight distribution of various modality data samples, thereby validating the model's ability to effectively identify target features and control false positives.

In the scenario shown in Fig. 9(b), the rainy season in the mountainous areas of Zhejiang brings humid, misty weather. The Grad-CAM heatmap reveals that the model focuses on the dense misty regions, yet it does not misclassify this common natural phenomenon as a fire. This indicates that the model not only detects the presence of smoke but also has a "comprehension" of natural smoke characteristics, effectively distinguishing between fire-related smoke and harmless environmental smoke under complex natural conditions. Another typical non-fire scenario, shown in Fig. 9(a), is a high-humidity lakeside area, which was deliberately designed as an adversarial example to further verify the model's cognitive depth and its ability to make judgments in special environments. The results demonstrate that while focusing on smoke regions, the model also considers water features. This sensitivity to environmental factors enables the model to make more accurate fire risk judgments in complex natural environments, showcasing high robustness and adaptability.

In the nighttime straw-burning scenario shown in Fig. 9(c), the

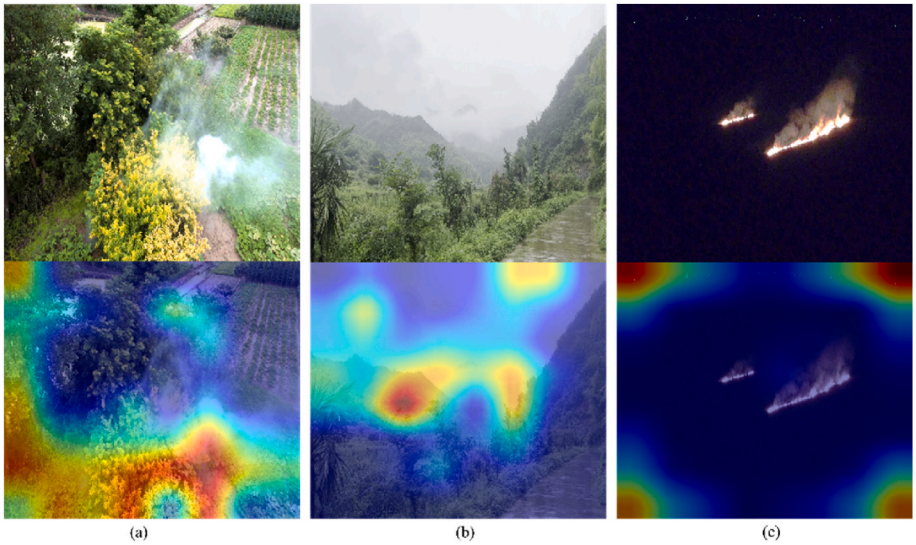

Fig. 9. Three typical samples for interpretable analysis: (a)Smoke beside the lake; (b)Misty rain in the mountains; (c)Burning straw at night.

model's attention is not focused on the smoke regions but instead shifts toward the illuminated residential areas in the four corners of the image. This attention shift exposes the limitations of relying solely on visual information in nighttime scenarios: in the absence of additional data support, the model tends to mistake bright non-fire light sources for fire. By incorporating multimodal information, the initial visual analysis can be corrected, offering an effective strategy for improvement.

5. Conclusions and future work

This study proposes a novel MM-SRENet model that integrates smoke images with fire risk data, effectively addressing the challenge that "visual smoke detection cannot directly assess fire risk" through a multimodal fusion strategy. The model incorporates a Multi-Scale Dilation Block (MSDBlock) with fine detail-capturing capability and additional detection heads in the object detection module. It also optimizes the original loss function with Selective Repulsion Loss to handle the diversity, blurriness, and occlusion of smoke in forest environments. The combination of extracted smoke scene patterns and specific fire risk factors enables MM-SRENet to demonstrate good adaptability in complex natural backgrounds. Experimental results show that MM-SRENet achieves a detection accuracy of 93.68 %, which is an 18.75 % improvement compared to traditional single-modality models. Specifically, the model's false positive rate is 7.09 %, below 10 %, meeting the false positive control requirements for practical deployment. In fire risk assessment, the model's risk prediction error is 0.15, indicating that abrupt changes in risk levels due to minor probabilistic fluctuations are limited. Furthermore, MM-SRENet exhibits minimal performance fluctuation across various environmental conditions, with an adaptability score of 0.91, demonstrating the model's good generalization ability in different environments.

In the future, our goal is to simulate an expert system to achieve a scientific and integrated forest fire risk assessment. We plan to further enrich the multimodal dataset by integrating multi-layered structured data such as historical fire records and geographic information to explore adaptive algorithms. With advancements in high-precision simulation technologies, generating forest fire data using virtual engines is expected to address the scarcity of real fire samples. By automating the extraction and annotation of key features, we aim to establish an integrated data generation pipeline that not only reduces human effort but also fully leverages the data-driven advantages of deep learning. We believe that, with continuous research progress, such a system will become a powerful tool for fire prevention decision-making, paving the way for a more scientific and intelligent approach to forest fire early warning.

CRediT authorship contribution statement

Peixian Jin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Pengle Cheng:** Project administration, Funding acquisition, Conceptualization. **Xiaodong Liu:** Project administration, Funding acquisition. **Ying Huang:** Supervision, Investigation.

Declaration of competing interest

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.

Acknowledgements

This study is supported by National Key R&D Program of China (2023YFC3006804) and National Natural Science Foundation of China (32171797),by Chunhui Project Foundation of the Education Department of China (HZKY20220026).

Data availability

Data will be made available on request.

References

Abid, F., 2021. A survey of machine learning algorithms based forest fires prediction and detection systems. Fire Technol. 57, 559–590. https://doi.org/10.1007/s10694-020-01056-z.

Aksoy, S., Haralick, R.M., n.d. Feature Normalization and Likelihood-Based Similarity Measures for Image Retrieval.

Baek, J., Alhindi, T.J., Jeong, Y.S., Jeong, M.K., Seo, S., Kang, J., Heo, Y., 2021. Intelligent multi-sensor detection system for monitoring indoor building fires. IEEE Sens. J. 21, 27982–27992. https://doi.org/10.1109/JSEN.2021.3124266.

Bhamra, J.K., Ramaprasad, S.A., Baldota, S., Luna, S., Zen, E., Ramachandra, R., Kim, H., Schmidt, C., Arends, C., Block, J., Perez, I., Crawl, D., Altintas, I., Cottrell, G.W., Nguyen, M.H., 2023. Multimodal wildland fire smoke detection. Remote Sens. 15. https://doi.org/10.3390/rs15112790.

Cao, Y., Yang, F., Tang, Q., Lu, X., 2019. An attention enhanced bidirectional LSTM for early forest fire smoke recognition. IEEE Access 7, 154732–154742. https://doi.org/10.1109/ACCESS.2019.2946712.

Chaturvedi, S., Khanna, P., Ojha, A., 2022. A survey on vision-based outdoor smoke detection techniques for environmental safety. ISPRS J. Photogrammetry Remote Sens. 185, 158–187. https://doi.org/10.1016/j.isprsjprs.2022.01.013.

Chen, S., Li, W., Cao, Y., Lu, X., 2022. Combining the convolution and transformer for classification of smoke-like scenes in Remote sensing images. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2022.3208120.

Chinembiri, T.S., Mutanga, O., Dube, T., 2024. A multi-source data approach to carbon stock prediction using Bayesian hierarchical geostatistical models in plantation forest ecosystems. GIsci Remote Sens 61. https://doi.org/10.1080/15481603.2024.2303868.

Coban, H.O., Bereket, H., 2020. Visibility analysis of fire lookout towers protecting the mediterranean forest ecosystems in Turkey. Sumar List 144 (5), 393–407. https://doi.org/10.31298/sl.144.7-8.

da Silva, D.Q., dos Santos, F.N., Sousa, A.J., Filipe, V., Boaventura-Cunha, J., 2021. Unimodal and multimodal perception for forest management: review and dataset. Computation 9. https://doi.org/10.3390/computation9120127.

de Venancio, P.V.A.B., Campos, R.J., Rezende, T.M., Lisboa, A.C., Barbosa, V.A., 2023. A hybrid method for fire detection based on spatial and temporal patterns. Neural Comput. Appl. 35, 9349–9361. https://doi.org/10.1007/s00521-023-08260-2.

Hu, Yaowen, Zhan, J., Zhou, G., Chen, A., Cai, W., Guo, K., Hu, Yahui, Li, L., 2022. Fast forest fire smoke detection using MVMNet. Knowl. Base Syst. 241. https://doi.org/10.1016/j.knosys.2022.108219.

Jing, T., Zeng, M., Meng, Q.-H., 2023. SmokePose: end-to-end smoke keypoint detection. IEEE Trans. Circ. Syst. Video Technol. 33, 5778–5789. https://doi.org/10.1109/TCSVT.2023.3258527.

Jones, M.W., Abatzoglou, J.T., Veraverbeke, S., Andela, N., Lasslop, G., Forkel, M., Smith, A.J.P., Burton, C., Betts, R.A., van der Werf, G.R., Sitch, S., Canadell, J.G., Santin, C., Kolden, C., Doerr, S.H., Le Quere, C., 2022. Global and regional trends and drivers of fire under climate change. Rev. Geophys. 60. https://doi.org/10.1029/2020RG000726.

Krippendorff, K., n.d. Computing Krippendorff's Alpha-Reliability.

Li, G., Cheng, P., Li, Y., Huang, Y., 2024. Lightweight wildfire smoke monitoring algorithm based on unmanned aerial vehicle vision. Signal Image Video Process 18, 7079–7091. https://doi.org/10.1007/s11760-024-03377-w.

Long, J.W., Drury, S.A., Evans, S.G., Maxwell, C.J., Scheller, R.M., 2022. Comparing smoke emissions and impacts under alternative forest management regimes. Ecol. Soc. 27. https://doi.org/10.5751/ES-13553-270426.

Ma, X., Dai, X., Bai, Y., Wang, Y., Fu, Y., n.d. Rewrite the Stars.

Park, M., Tran, D.Q., Bak, J., Park, S., 2022. Advanced wildfire detection using generative adversarial network-based augmented datasets and weakly supervised object localization. Int. J. Appl. Earth Obs. Geoinf. 114. https://doi.org/10.1016/j.jag.2022.103052.

Phan, T.C., Nguyen, T.T., Hoang, T.D., Nguyen, Q.V.H., Jo, J., 2020. Multi-scale bushfire detection from multi-modal streams of Remote sensing data. IEEE Access 8, 228496–228513. https://doi.org/10.1109/ACCESS.2020.3046649.

Prema, C.E., Suresh, S., Krishnan, M.N., Leema, N., 2022. A novel efficient video smoke detection algorithm using Co-occurrence of local binary pattern variants. Fire Technol. 58, 3139–3165. https://doi.org/10.1007/s10694-022-01306-2.

Rui, X., Li, Ziqiang, Zhang, X., Li, Ziyang, Song, W., 2023. A RGB-Thermal based adaptive modality learning network for day-night wildfire identification. Int. J. Appl. Earth Obs. Geoinf. 125. https://doi.org/10.1016/j.jag.2023.103554.

Sathishkumar, V.E., Cho, J., Subramanian, M., Naren, O.S., 2023. Forest fire and smoke detection using deep learning-based learning without forgetting. FIRE ECOLOGY 19. https://doi.org/10.1186/s42408-022-00165-0.

Seidl, R., Thom, D., Kautz, M., Martin-Benito, D., Peltoniemi, M., Vacchiano, G., Wild, J., Ascoli, D., Petr, M., Honkaniemi, J., Lexer, M.J., Trotsiuk, V., Mairota, P., Svoboda, M., Fabrika, M., Nagel, T.A., Reyer, C.P.O., 2017. Forest disturbances under climate change. Nat. Clim. Change 7, 395–402. https://doi.org/10.1038/NCLIMATE3303.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., n.d. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.

Shaharuddin, S., Maulud, K.N.A., Rahman, S.A.F.S.A., Ani, A.I.C., Pradhan, B., 2023. The role of IoT sensor in smart building context for indoor fire hazard scenario: a

systematic review of interdisciplinary articles. INTERNET OF THINGS 22. https://doi.org/10.1016/j.iot.2023.100803.
Vikram, R., Sinha, D., De, D., Das, A.K., 2020. EEFFL: energy efficient data forwarding for forest fire detection using localization technique in wireless sensor network. Wirel. Netw. 26, 5177–5205. https://doi.org/10.1007/s11276-020-02393-1.
Wang, L., Li, Haiyan, Siewe, F., Ming, W., Li, Hongsong, 2024. Forest fire detection utilizing ghost Swin transformer with attention and auxiliary geometric loss. Digit. Signal Process. 154. https://doi.org/10.1016/j.dsp.2024.104662.
Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C., 2017. Repulsion Loss: Detecting Pedestrians in a Crowd.
Wu, L., Chen, L., Hao, X., 2021. Multi-sensor data fusion algorithm for indoor fire early warning based on BP neural network. Information 12. https://doi.org/10.3390/info12020059.
Wu, S., Zhang, X., Liu, R., Li, B., 2023. A dataset for fire and smoke object detection. Multimed. Tool. Appl. 82, 6707–6726. https://doi.org/10.1007/s11042-022-13580-x.
Yin, D.X., Cheng, P., Huang, Y., 2024. YOLO-EPF: multi-scale smoke detection with enhanced pool former and multiple receptive fields. Digit. Signal Process.: A Review Journal 149. https://doi.org/10.1016/j.dsp.2024.104511.
Yuan, C., Liu, Z., Zhang, Y., 2017. Aerial images-based forest fire detection for firefighting using optical Remote sensing techniques and unmanned aerial vehicles. J. Intell. Rob. Syst. 88, 635–654. https://doi.org/10.1007/s10846-016-0464-7.
Yuan, F., Shi, Y., Zhang, L., Fang, Y., 2023. A cross-scale mixed attention network for smoke segmentation. Digit. Signal Process. 134. https://doi.org/10.1016/j.dsp.2023.103924.
Yuan, F., Zhang, L., Xia, X., Wan, B., Huang, Q., Li, X., 2019. Deep smoke segmentation. Neurocomputing 357, 248–260. https://doi.org/10.1016/j.neucom.2019.05.011.
Zhao, Y., Zheng, Q., Zhu, P., Zhang, X., Ma, W., 2024. TUFusion: a transformer-based universal fusion algorithm for multimodal images. IEEE Trans. Circ. Syst. Video Technol. 34, 1712–1725. https://doi.org/10.1109/TCSVT.2023.3296745.