forests

MDPI

Article

Early Smoke Recognition Algorithm for Forest Fires

Yue Wang, Yan Piao *, Qi Wang, Haowen Wang, Nan Qi and Hao Zhang

School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China; 2021200103@mails.cust.edu.cn (Y.W.); 2021200102@mails.cust.edu.cn (Q.W.); wanghaowen@mails.cust.edu.cn (H.W.); 2021200101@mails.cust.edu.cn (N.Q.); zhanghao@mails.cust.edu.cn (H.Z.)

* Correspondence: piaoyan@cust.edu.cn; Tel.: +86-180-8863-0051

Abstract: Forest fires require rapid and precise early smoke detection to minimize damage. This study focuses on employing smoke recognition methods for early warning systems in forest fire detection, identifying smoke as the primary indicator. A significant hurdle lies in the absence of a large-scale dataset for real-world early forest fire smoke detection. Early smoke videos present characteristics such as smoke plumes being small, slow-moving, and/or semi-transparent in color, and include images where there is background interference, posing critical challenges for practical recognition algorithms. To address these issues, this paper introduces a real-world early smoke monitoring video dataset as a foundational resource. The proposed 4D attention-based motion target enhancement network includes an important frame sorting module which adaptively selects essential frame sequences to improve the detection of slow-moving smoke targets. Additionally, a 4D attention-based motion target enhancement module is introduced to mitigate interference from smoke-like objects and enhance recognition of light smoke during the initial stages. Moreover, a high-resolution multi-scale fusion module is presented, incorporating a small target recognition layer to enhance the network's ability to detect small smoke targets. This research represents a significant advancement in early smoke detection for forest fire surveillance, with practical implications for enhancing fire management.

Keywords: early smoke recognition; forest fires; motion target enhancement network

Citation: Wang, Y.; Piao, Y.; Wang, Q.; Wang, H.; Qi, N.; Zhang, H. Early Smoke Recognition Algorithm for Forest Fires. *Forests* **2024**, *15*, 1082. https://doi.org/10.3390/f15071082

Academic Editor: Luis A. Ruiz

Received: 23 May 2024
Revised: 17 June 2024
Accepted: 20 June 2024
Published: 22 June 2024

1. Introduction

Forest fires are a major environmental disaster facing forest resources. Once a forest fire occurs, it can quickly get out of control, and then require enormous effort, time, and resources to extinguish. Additionally, forest fires and related combustion processes release large amounts of pollutants into the atmosphere, including CO_2, CO, CH_4, BC, VOCs, $PM_{2.5}$, and PM_{10}. CO_2 and CH_4 directly increase the concentration of greenhouse gases in the atmosphere, while CO can oxidize to CO_2 in the atmosphere. BC exacerbates global warming by absorbing solar radiation and altering the albedo of ice and snow. VOCs, $PM_{2.5}$, and PM_{10} reduce air quality and worsen air pollution. Consequently, these pollutants have significant impacts on environmental pollution and global warming. In recent years, numerous studies have focused on various pollutants and their effects on the environment and climate. Gürbüz et al. [1] proposed a computational approach for predicting the environmental impact of pollutants emitted from the transportation sector. Meanwhile, Ekici et al. [2] introduced a novel transparent computational method for analyzing the quantity of aviation pollutants. Smoke is always the first sign seen when a forest fire occurs. Therefore, early smoke recognition in forest fires is significant for early warning. Early smoke detection methods mainly rely on traditional image processing methods [3–6] and deep learning methods [7–16].

In addition to examining the spatial properties of images, traditional image analysis techniques extensively employ transform domain analysis, with the wavelet transform

being a notable example. Barmpoutis et al. [3] presented a real-time smoke detection algorithm for video streams, incorporating background subtraction, color analysis, spatial energy analysis, spatiotemporal analysis, histogram analysis, and dynamic texture analysis. Dimitropoulos et al. [4] introduced an algorithm for recognizing smoke in videos using descriptors derived from linear dynamical systems. Islam et al. [5] proposed a smoke detection method that combines mixture smoke segmentation with an efficient dynamic smoke symmetry model. Recently, Wu et al. [6] developed a patchwork dictionary learning approach for early smoke detection in forest fires. However, these methods face challenges such as computational complexity, uncertainty, and insufficient precision in feature selection, resulting in weak adaptability and high computational resource requirements.

With the development and progress of deep learning, detection algorithms based on deep learning have become more efficient, reducing hardware costs and eliminating the need for manual feature extraction compared to traditional algorithms. Hu et al. [7] presented a method for video smoke recognition using a spatio-temporal CNN. Aslan et al. [8] proposed a motion-based geometric image transformation and DCGAN for wildfire smoke recognition. Yang et al. [9] introduced a smoke detection technique utilizing the DenseNet neural network architecture. Hsu et al. [10] introduced the pioneering RISE dataset, a comprehensive video dataset designed explicitly for identifying industrial smoke releases, using the Inception-v1 I3D neural network architecture for smoke recognition, which is capable of distinguishing smoke from water vapor. Shi et al. [11] employed a compact deep convolutional neural network for recognizing smoke in videos. Tao et al. [12] proposed a deformable convolutional enhancement network based on semantic correlation multidirectional interaction for forest smoke recognition. Jiang et al. [13] introduced an attention mechanism in the EfficientNet network for smoke recognition during straw burning. Cao et al. [14] developed a smoke source recognition and prediction method based on the enhanced feature foreground network for wildfire smoke recognition. Li et al. [15] proposed a 3D parallel fully convolutional network architecture for wildfire smoke recognition. Zhu et al. [16] proposed a 3D convolutional encoder–decoder network architecture for smoke recognition. The existing algorithms have only achieved satisfactory performance on synthetic, self-made, small-scale, and non-forest datasets. This is primarily because the smoke images in these datasets feature simple backgrounds, clearly visible smoke movement, large target sizes, and dense colors. However, these conditions differ significantly from those in forest environments. In forest scenarios, smoke is often small, light, slow-moving, and surrounded by complex and variable backgrounds. Moreover, current smoke detection algorithms lack modules to handle these challenging conditions. Consequently, existing algorithms cannot be directly applied to forest environments for detecting small, light, and slow-moving smoke amidst complex backgrounds.

To address the above issues, this paper first integrates a forest fire early smoke surveillance video dataset in a real scenario with all the early smoke features, laying a good foundation for future practical applications. Then, this paper introduces a 4D-MENet (4D attention-based motion target enhancement network) for early smoke recognition in forest fire monitoring videos. The 4D-MENet contains three new modules: FS (important frame sorting module), 4D-ME (4D attention-based motion target enhancement module), and HFM (high-resolution multi-scale fusion module). Our contribution is fourfold:

1. To detect smoke targets in complex backgrounds and improve the feature representation of light smoke targets, a 4D-ME module is proposed which enhances the neural network's attention to temporal, spatial, and color features of smoke, improves the recall of smoke recognition, and reduces the false alarm rate.
2. To detect slow-moving smoke targets, an FS module is proposed to adaptively extract significant frame sequences from the input image sequence, facilitating the subsequent network to extract motion features of slow-moving smoke.
3. To detect small smoke targets, an HFM module is proposed to add a small target recognition layer which fuses shallow small target feature information in the high-

level feature map, thus, enhancing the high-level feature map's small smoke target recognition capability.

4. In this paper, we integrate a large-scale video dataset of forest fire early smoke in real scenarios containing various challenging features of early smoke. To save human, material, and financial costs, this paper annotates them with categorical labels for subsequent supervised learning of neural networks, including 2450 smoke sequences and 3800 non-smoke sequences, laying a good foundation for future practical applications.

2. Materials and Methods

2.1. Problems and Motivations

This section analyzes the current problem of forest fire early smoke detection and the solutions to different issues to make the innovation of this paper more straightforward.

Problem 1: Repeatedly extracting the same features is not conducive to model training and convergence. This issue arises mainly due to the slow movement of smoke at the beginning of forest fire surveillance videos. As a result, several consecutive frames in the video are nearly identical, with minimal movement changes, causing the network model to repeatedly extract the same features, hindering practical training and convergence. Typically, the intervals between important frames vary across different video sequences. Therefore, most current smoke detection methods employ random sampling of other video parts [14], allowing the network model to extract diverse features and enhance the representation of smoke characteristics. In this paper, we utilize an adaptive important frame sorting module to preprocess the input image sequences, enabling the model to automatically extract the crucial frames from each input sequence, thereby enhancing feature representation.

Problem 2: Existing forest early smoke detection algorithms have low accuracy and high false alarm rates. This issue primarily stems from the scarcity of labeled early real fire smoke video datasets, leading to non-robust extraction of smoke features and impacting the performance of early smoke detection algorithms. Since creating a large number of pixel-level labels requires substantial human, material, and financial resources, current algorithms mostly rely on manually labeling a small number of pixel-level labels and then employing GAN networks to generate datasets [8] or directly synthesizing datasets based on smoke imaging principles [12] to mitigate the challenges above. However, these approaches result in complex models, with only marginal improvements in detection performance, rendering them unsuitable for direct application in real-world scenarios. In this study, we adopt manually labeled smoke and non-smoke video classification labels for supervised training, enhancing local semantic relevance compared to pixel-level labels. Additionally, we incorporate a motion enhancement network to prioritize the motion region of the image. Moreover, considering the characteristics of early smoke videos from forest fires—such as small imaging targets, semi-transparent color, and interference from similar smoke targets in the background—existing smoke detection algorithms face significant challenges. Most algorithms employ spatial attention, channel attention, self-attention, and multi-scale fusion methods to mitigate these challenges. However, they often overlook the temporal characteristics of smoke plume movement, resulting in complex models with minimal improvements in detection performance. To address these issues and enhance the detection ability of light smoke, we introduce non-smoke labels that are similar to smoke in the training set. We also utilize a moving target enhancement network based on the 4D attention mechanism to improve the network's smoke feature extraction capability. Furthermore, to enhance the representation of small smoke target features, we integrate a high-resolution multi-scale feature fusion module and add a small target detection layer to the network.

2.2. Datasets

Experiments are conducted on two datasets to demonstrate the proposed algorithm's performance advantages. The first dataset is a natural forest fire smoke video's dataset

integrated into this paper, which is named the Forest Smoke dataset. Since this paper studies early smoke recognition in forest fires, selecting a dataset with challenging features such as small smoke imaging target, semi-transparent color, slow movement, and interference with smoke-like targets is necessary. The dataset proposed in this study consists of two categories comprising 6250 videos, with 2450 videos containing smoke and 3800 videos without smoke. All videos have a resolution of 1920 × 1080 and are in mp4 format. The dataset was collected through various methods: downloading from internet websites and search engines, integrating currently available public datasets, extracting from movie clips, and manually capturing images using different models of smartphone and digital camera. These approaches ensure that the dataset is both challenging and diverse. Among these, 4375 videos are used for training, 625 for validation, and 1250 for testing the experimental outcomes. The video images in the training set are not directly used for training and need to be labeled manually with classification labels. The annotation method employed for the dataset proposed in this study includes data cleaning, tool configuration, personnel training, task assignment, interim review, and final review. Quality assurance measures comprise multiple rounds of annotation, review mechanisms, automated quality checks, and ongoing improvement efforts.

The second dataset is the industrial smoke emission dataset, RISE, proposed by Hsu et al. [10]. This dataset contains 12,567 time-lapse video clips from three sites covering different scenes across four seasons of daytime over two years, which have been practically applied. RISE splits the dataset into the training set, validation set, and test set in six ways ($S_0, S_1, S_2, S_3, S_4, S_5$); S_3 is divided based on the time sequence, and the other five are split based on the camera view. All videos have a resolution of 180 × 180 and are in mp4 format.

The image sequences of smoke and non-smoke for the two datasets are shown in Figure 1, where the non-smoke image sequence in the Forest Smoke dataset was chosen from the smoke-like image sequence, which often occurs in forests and is remarkably similar to smoke, and the non-smoke image sequence in the RISE dataset was selected from the water vapor image sequence, which often occurs in industry and is quite similar to smoke.

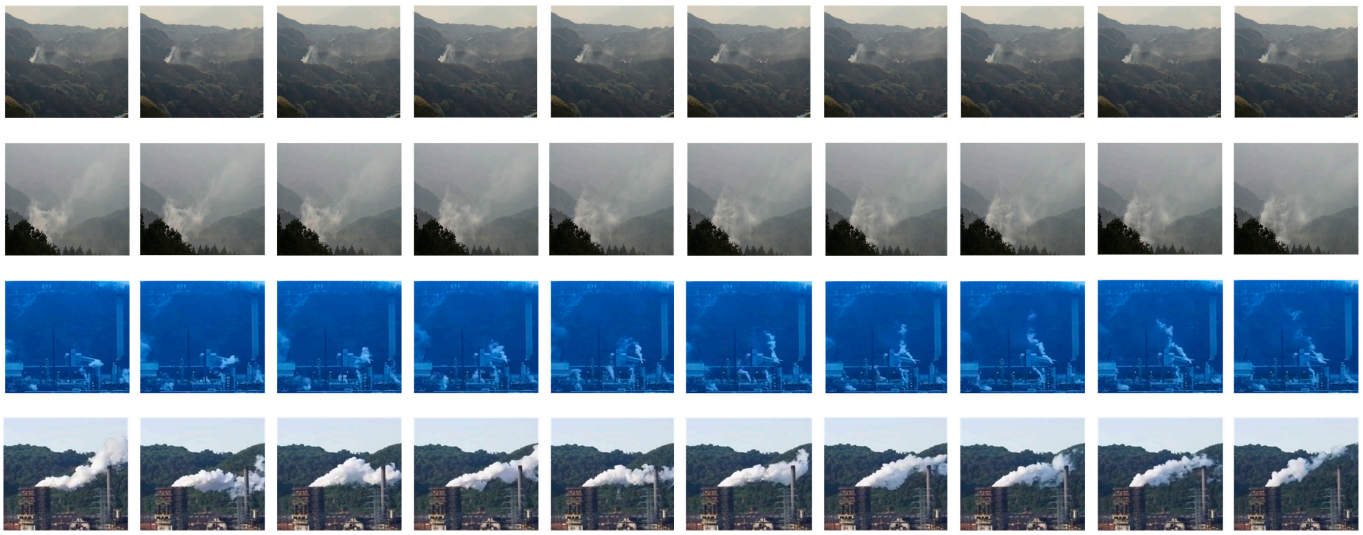

Figure 1. The first row is smoke from the Forest Smoke dataset, the second row is non-smoke (fog) from the Forest Smoke dataset, the third row is smoke from the RISE dataset, and the fourth row is water vapor from the RISE dataset.

2.3. 4D Attention-Based Motion Target Enhancement Network

2.3.1. Proposed Network Architecture

Figure 2 shows the entire network framework of our 4D-MENet. The network architecture mainly includes FS, 4D-ME, and HFM modules.

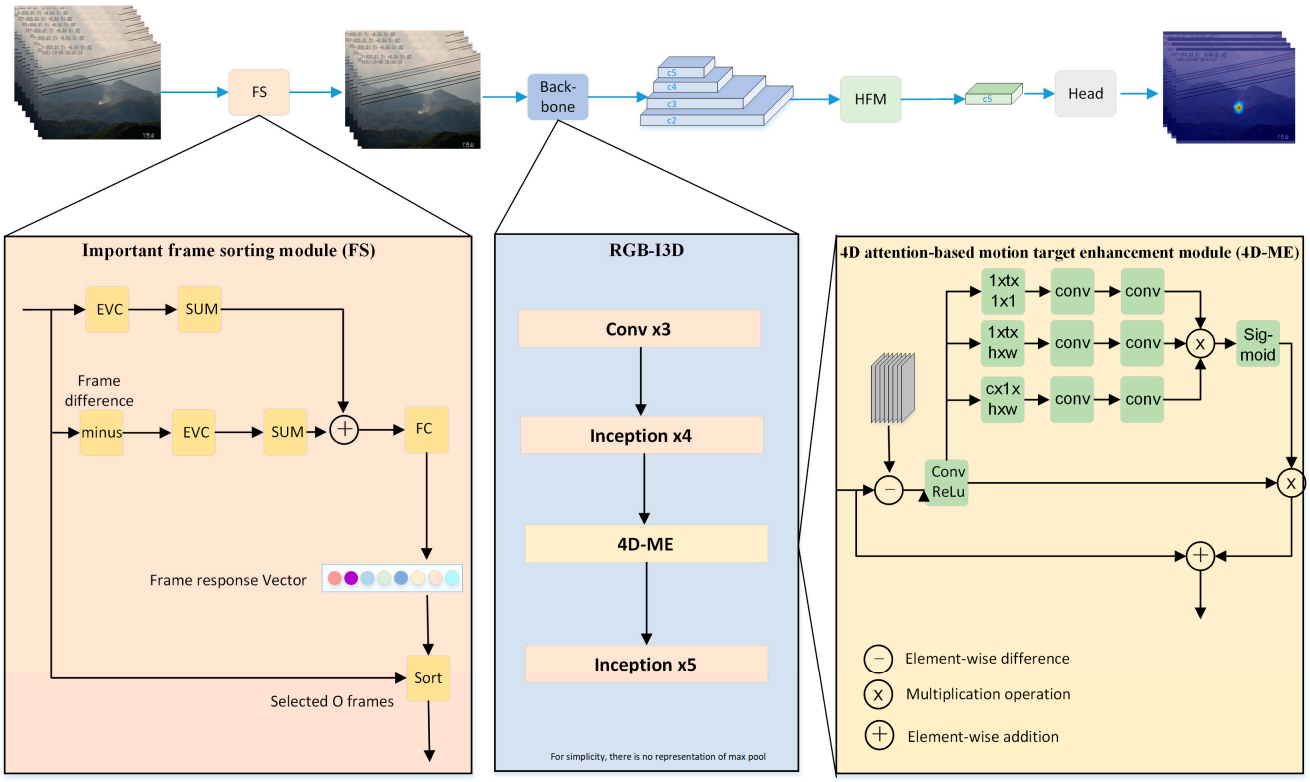

Figure 2. The whole network architecture of the proposed 4D-MENet.

As shown in Figure 2, the network initially inputs an image sequence. First, the important frame sequence of the input image is adaptively extracted through the FS. Second, the important frame sequence is put into the backbone to extract robust multi-scale features through the 4D-ME. Then, the feature maps with resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image are sent to the HFM for top-to-bottom and bottom-to-top dual-path cross-fertilization. The highest-level semantic features and the lowest-level spatial features are integrated into the middle feature map to enhance the features' ability to be represented. Finally, the highest-level feature map is sent to the prediction network for prediction.

The FS removes redundant information from the input image sequence. It extracts important frame sequences, which lays a good foundation for subsequent neural network motion feature extraction. This paper adds a 4D-ME to the backbone, which solves the local semantic correlation of the labels in the classification annotation dataset and makes the network pay more attention to the moving targets, thereby distinguishing smoke-like targets. The HFM introduces the high-resolution feature map of 1/4 of the original map to increase the image recognition layer of small targets. At the same time, the multi-path enhancement of feature maps with different resolutions can improve the semantic and spatial properties of high-level feature maps, which lays an excellent foundation for the subsequent prediction of the classification accuracy of the network.

2.3.2. Important Frame Sorting Module

The slow movement of smoke at the beginning of the forest fire surveillance video, resulting in several consecutive frames of the surveillance video being the same, with almost no movement changes, makes the network model constantly extract the same features and is not conducive to model training and convergence. Traditional image preprocessing usually uses random extraction of important frames to reduce the redundant information of the input image sequence, losing a lot of important information, so we use FS to extract the important frame sequence of the input image, as shown in Figure 2. FS mainly contains

two parts: the first part extracts the feature information of each frame, and the second part extracts the feature information of the difference between two neighboring frames.

Assume that the input image sequence is $P = \{P_1, P_2, \cdots P_I\}$, and the output image sequence is $T = \{T_1, T_2, \cdots T_O\}$, the entire procedure is shown below:

$$F_i = \|EVC(P_i)\|, i = 1, 2, \cdots, I. \tag{1}$$

$$D_i = \|EVC(P_{i+1} - P_i)\|, i = 1, 2, \cdots, I. \tag{2}$$

$$R = FC(F + D). \tag{3}$$

where non-existent P_{L+1} in the above formula is set to P_{L-2} so that F and D have the same dimension. F_i represents the feature information of each frame, D_i represents the feature information of the difference between adjacent frames, FC represents the full convolution, EVC [17] is the explicit visual center module, which can extract not only the contextual information of the image but also the information of the local corners of the smoke, and R represents the response vector of the importance of each frame.

In this paper, the following formula is used to select the important O frames:

$$T_o = P_{ind(S(o))}, o = 1, 2, \cdots O; O < I \tag{4}$$

$$S = sort(R) \tag{5}$$

where $ind(S(o))$ denotes the index of $S(o)$ in R, T_O denotes the adaptively extracted sequence of important frames, and S denotes the sorting of the significance response vectors.

To maximize accuracy and minimize false alarms, we should reduce redundant information in the input sequences so the setting in the important frames sorting module is important. It can be adjusted appropriately according to your experimental data; see Section 3.4.1 for details.

2.3.3. 4D Attention-Based Motion Target Enhancement Module

The backbone architecture RGB-I3D [18] used in this article is input as an RGB frame, which is expanded on inception_v1. To effectively retain the smoke movement time information, the kernel corresponding to the first two max pooling layers is $1 \times 3 \times 3$, and the stride is $1 \times 2 \times 2$; the subsequent kernel and stride are normal expansion.

To solve the problem of semantic correlation of labels in the Forest Smoke dataset, this paper adds a motion target enhancement module to the backbone to guide the model to learn discriminative features from the complex background. On the other hand, to distinguish the interference of smoke-like targets and improve light smoke recognition capabilities, this paper superimposes a 4D attention mechanism on the moving target enhancement network to make the network extract discriminative features, and the specific network structure is shown in Figure 2.

The motion target enhancement module proposed in this paper is to make a difference between each frame and the simulated background frame to obtain the motion region of the image and then superimpose the temporal, spatial, and channel attention mechanism, that is, the 4D attention mechanism so that the network can extract discriminative features which can distinguish the interference of smoke-like targets and improve light smoke recognition capabilities.

Given the important frame sequence obtained from FS $f_{in} \in \mathbb{R}^{T \times C \times H \times W}$, through a series of RGB-I3D convolutions to obtain the middle layer features $f_{mid} \in \mathbb{R}^{T \times C \times H \times W}$, where T represents time, C denotes the number of channels, and H and W represent the spatial dimensions of height and width, the connection among the middle layer feature f_{mid} and the input video f_{in} can be mathematically described by the following equation:

$$f_{in} = conv(f_{in}, W_\theta) \tag{6}$$

where conv represents the RGB-I3D convolutional block and W_θ represents the trainable parameters of the RGB-I3D convolutional network. After that, all the frames in the middle layer are summed and averaged to simulate the background image, and finally, the background image is subtracted from each frame to obtain the motion target enhancement area, which the following equation can mathematically describe:

$$f_{en} = f_{\mathrm{mid}} - \frac{1}{T}(\Sigma_t f_{\mathrm{mid}}) \tag{7}$$

Since direct subtraction will lead to problems such as model convergence difficulties and data overflow, it is essential to carry out convolution and nonlinear operations on the motion target enhancement area to enhance the representation ability of features. Specifically, it can be mathematically described by the following equation:

$$f_{en} = conv\left(\delta\left(f_{fore}\right), W_\varphi\right) \tag{8}$$

where $\delta(.)$ represents ReLU function and W_φ is a parameter that can be learned by convolution.

At present, the main attention mechanisms are SA [19], CA [20], and CBAM [21]. To make the convolutional neural network pay more attention to the motion region in the forest surveillance video to extract the discriminative smoke features, this paper superimposes the 4D attention mechanism on the extracted motion target enhancement region. The direct learning of the 4D attention mechanism is costly because it will contain a large number of parameters. Hence, this paper decomposes the 4D attention mechanism into three low-dimensional attention mechanisms for learning to reduce complexity and the number of parameters. Specifically, it can be mathematically described by the following equation:

$$S_a = conv\left(\mathrm{ReLU}\left(conv\left(\frac{1}{T}\sum_{t=1}^{T} x_{1:C,t,1:H,1:W}\right)\right)\right) \tag{9}$$

$$C_a = conv\left(\mathrm{ReLU}\left(conv\left(\frac{1}{T\times H\times W}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{1:C,t,h,w}\right)\right)\right) \tag{10}$$

$$T_a = conv\left(\mathrm{ReLU}\left(conv\left(\frac{1}{C\times H\times W}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{c,1:T,h,w}\right)\right)\right) \tag{11}$$

$$f_{out} = f_{\min} + Sigmoid(S_a \times C_a \times T_a) f_{en} \tag{12}$$

$S_a \in \mathbb{R}^{(1\times1\times H\times W)}, C_a \in \mathbb{R}^{(C\times1\times1\times1)}, T_a \in \mathbb{R}^{(1\times T\times1\times1)}, f_{en} \in \mathbb{R}^{(T\times C\times H\times W)}, f_{out} \in \mathbb{R}^{(T\times C\times H\times W)}$ represent the spatial attention coefficient, channel attention coefficient, temporal attention coefficient, moving target enhanced feature map, and output feature map, respectively. S_a first aggregates temporal information through global average pooling with respect to time and then calculates spatial attention by changing the channel dimension to 1 through 2 convolutions. C_a first aggregates temporal and spatial information through global average pooling with respect to time and space, then the channel attention is calculated through 2 convolutions. T_a first aggregates channel and spatial information through global average pooling with respect to channel and space, then calculates the temporal attention through 2 convolutions, and finally multiplies the product of attention through the sigmoid function. The quantization is 0–1 and is superposed to the moving target enhancement module to calculate the network output feature map.

2.3.4. High-Resolution Multi-Scale Fusion Module

Since low-scale feature maps have better spatial information and poorer semantic information, and high-scale feature maps have better semantic information and more deficient spatial information, the current neural network cross-fuses different scale features to express feature spatial information and semantic information [22–26]. However, since the smoke in the forest fire smoke monitoring videos only accounts for a small part of

the whole image, this paper introduces a feature map with 1/4 resolution of the original image into the multi-scale feature fusion network, which makes the network increase the small target image recognition layer. On the other hand, to better integrate high-level semantic and low-level spatial information, this paper introduces the fusion of high-level and low-level feature maps on the intermediate feature maps. Therefore, the backbone calculates the feature maps with resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image, which is then sent to the HFM for top-to-bottom and bottom-to-top dual-path cross-fertilization. The highest-level semantic and lowest-level spatial features are integrated into the middle feature to enhance the ability to represent feature maps. The high-resolution multi-scale feature fusion module proposed in this article is shown in Figure 3. The network architecture includes top-to-bottom and bottom-to-top paths and the fusion of a single highest-level feature map and a single lowest-level feature map. Finally, the weighted fusion method improves the feature expression ability more comprehensively. The specific formula is as follows:

$$C_4^{td} = conv\left(\frac{w_1 \cdot C_4^{in} + w_2 \cdot \mathrm{Resize}\left(C_5^{in}\right)}{w_1 + w_2 + \varepsilon}\right) \tag{13}$$

$$C_4^{out} = conv\left(\frac{w_1' \cdot C_4^{in} + w_2' \cdot C_4^{td} + w_3' \cdot \mathrm{Resize}\left(C_3^{out}\right) + w_4' \cdot \mathrm{Resize}\left(C_2^{in}\right)}{w_1' + w_2' + w_3' + w_4' + \varepsilon}\right) \tag{14}$$

where w_1, w_2, w_1', w_2', w_3', and w_4' are network learnable parameters, ε is the minimum value to prevent the denominator from being 0, C_4^{td} is the middle layer feature of the network, C_5^{in}, C_4^{in}, and C_2^{in} are network input features, and C_3^{out} and C_4^{out} are network output features.

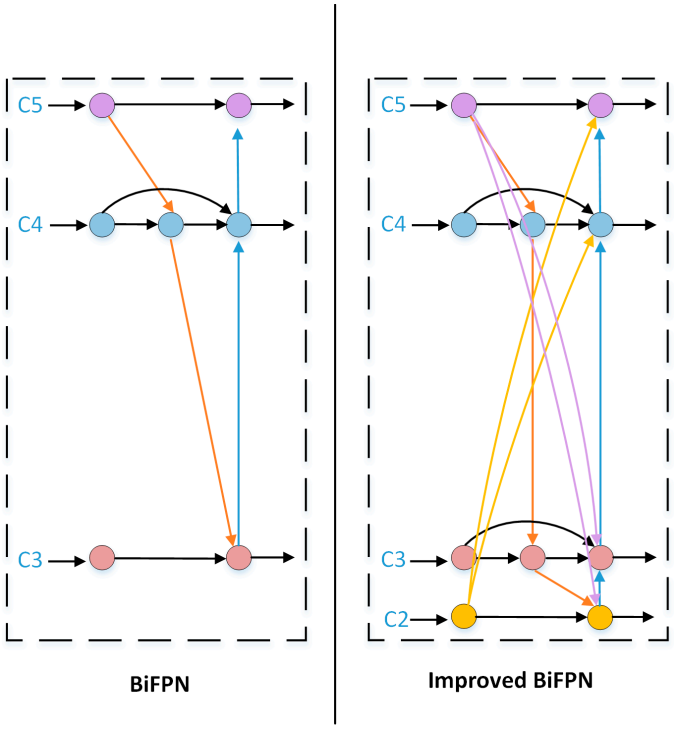

Figure 3. HFM network structure. In the figure, C2, C3, C4, and C5 represent feature maps computed by the backbone with resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image, respectively. The yellow arrow in the figure represents the low-level feature data flow, the purple arrow represents the high-level feature data flow, the orange arrow represents the top-down data flow, and the blue arrow represents the bottom-up data flow.

The repeated use of multiple HFM can make the extracted features richer, but too much repetition will not increase the performance of the algorithm but will increase the burden of computation for the model; the specific repetition of the parameter settings needs to be adjusted according to the results of the experiments; see Section 3.4.3 for details.

3. Results

This chapter first introduces experiment details and evaluation criterion. It then compares the current state-of-the-art smoke detection algorithms from both qualitative and quantitative perspectives. Finally, it discusses ablation experiments and practical applications.

3.1. Experiments Details

The experimental setup in this paper utilizes a personal desktop computer as the platform for conducting the experiments; the practical environment is an Ubuntu system, the processor is AMD Ryzen 9 5900X 12-Core Processor, and the graphics cards are multiple NVIDIA GeForce RTX 3090 GPU and PyTorch framework. The AMD Ryzen 9 5900X 12-Core Processor is manufactured by AMD, located in Santa Clara, CA, USA. The NVIDIA GeForce RTX 3090 GPU is manufactured by GAINWARD, located in Taiwan, China. In this study, we fine-tuned the learning rate, batch size, optimizer selection, regularization parameters, and network structure parameters. We initially used default parameters to initialize the model for the learning rate and conducted a broad range test to identify a rough range, gradually increasing from 0.0001 to 0.1. We refined the search within this range and recorded the model performance for each learning rate, ultimately selecting the best performance. We fixed other parameters for batch size at the initially determined optimal learning rate and tested different batch sizes, such as 4, 16, 32, and 64. We recorded the training time and model performance and selected the batch size that provided the best performance with a reasonable training time. In optimizing the optimizer, we compared the performance and training time of the model using the default parameters of SGD, Adam, and RMSprop. We chose the optimizer that performed best on the validation set. We set an initial range for regularization parameters and conducted cross-validation under different regularization coefficients, selecting the coefficient that minimized the validation set loss. Finally, we performed experiments with different convolutional kernel sizes and layer combinations for the network architecture parameters under optimal learning rate, batch size, and operation timing. The convolutional kernels of sizes $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $5 \times 5 \times 5$ were primarily tested in the inception modules of the RGB-I3D network. Additionally, different layers were experimented with by incorporating 4D-ME at various positions within the backbone network. By balancing performance and resource consumption, we selected the best network structure. Throughout the training phase, the learning rate at the beginning is set to 0.1, the milestones are (500, 1500), the attenuation weight is 10^{-6}, and the video shape of the input network is [40, 3, 36, 224, 224], corresponding to the batch size, the number of input channels, number of frames, and image height and width.

3.2. Evaluation Criterion

The evaluation metrics employed in this experiment include recall (R), the harmonic mean of precision and recall (F1-score), and false positive rate (FPR), which the following equation can mathematically describe:

$$R = TP/(TP + FN) \tag{15}$$

$$F1\text{-score} = 2TP/(2TP + FP + FN) \tag{16}$$

$$FAR = FP/(FP + TN) \tag{17}$$

where TP represents true positives, the number of instances correctly classified as the positive class. FN represents false negatives, the number of cases incorrectly classified as harmful. FP represents false positives, the number of instances incorrectly classified as the positive class. TN represents true negatives, the number of cases correctly classified as harmful.

3.3. Comparative Experiments

3.3.1. Quantitative Analysis

In this paper, we first compare 12 state-of-the-art smoke recognition algorithms on the integrated Forest Smoke dataset, and the specific comparison results are shown in Table 1. As shown in Table 1, the proposed algorithm significantly improves detection performance, with a recall (R) 1.17% higher and a false alarm rate (FAR) 1.61% lower than the second-best algorithm, EFFNet. This improvement is primarily due to EFFNet's approach of randomly extracting different frames from the input video sequence as keyframes, which introduces image redundancy and loss of important information. In contrast, our algorithm employs an important frame sorting module to adaptively extract critical frames from the input video sequence, enriching the features extracted by the neural network and enhancing the representation of smoke features, thereby improving the network's recall rate. Additionally, EFFNet uses a 2D attention mechanism to enhance smoke targets, while our algorithm utilizes a 4D attention mechanism. This enhances the distinction between smoke features and similar objects, improves the feature extraction capability for light smoke, and reduces the false alarm rate. Then, to prove the proposed algorithm's advantages, this paper compares it with 12 advanced smoke recognition algorithms on the RISE dataset. The specific comparison results are shown in Table 2. As shown in Table 2, the proposed algorithm in this paper achieves the best F1-score on S_0, S_1, and S_3. Especially on S_3, the F1-score is improved by 3% compared to the second-best method (I3D), mainly because the increased 4D-ME module adds the temporal attention mechanism. It, thus, has a more significant impact on the classification results of time series.

Table 1. Results of different smoke recognition methods on Forest Smoke.

Different Methods	R (%)	FAR (%)	F1-Score
SAN-SD [13]	93.67	7.2	0.9281
EFFNet [14]	94.45	5.73	0.9375
VSSNet [16]	94.13	5.68	0.9315
3D-PFCN [15]	94.04	5.48	0.9343
I3D [18]	93.56	5.55	0.9314
CNN-LSTM [27]	93.23	8.12	0.9261
DCNN [28]	92.16	8.12	0.9149
MobileNetv2 [11]	92.15	8.32	0.9151
DenseNet [9]	91.27	7.68	0.9112
DCGAN [8]	93.27	7.22	0.9258
C3D [29]	93.73	5.88	0.9305
MOG-CNN [30]	92.16	8.16	0.9149
4D-MENet (ours)	**95.62**	**4.12**	**0.9486**

Table 3 shows some of the parameters of the different models, and it can be seen that our algorithm trades only a small increase in parameters for a significant performance improvement.

Table 2. Results of different smoke recognition methods on RISE.

Different Methods	F1-Score					
	S_0	S_1	S_2	S_3	S_4	S_5
Flow-SVM [18]	0.42	0.59	0.47	0.63	0.52	0.47
Flow-I3D [1] [18]	0.55	0.58	0.51	0.68	0.65	0.50
SVM [18]	0.57	0.70	0.67	0.67	0.57	0.53
I3D [1] [18]	0.80	0.84	0.82	0.87	0.82	0.75
I3D [1]-ND [2] [18]	0.76	0.79	0.81	0.86	0.76	0.68
I3D [1]-FP [3] [18]	0.76	0.81	0.82	0.87	0.81	0.71
I3D [1]-TSM [4] [31]	0.81	0.84	0.82	0.87	0.80	0.74
I3D [1]-LSTM [5] [18]	0.80	0.84	0.82	0.85	**0.83**	0.74
I3D [1]-NL [6] [32]	0.81	0.84	0.83	0.87	0.81	0.74
I3D [1]-TC [7] [33]	0.81	0.84	0.84	0.87	0.81	0.77
CNN-NonFFM [14]	0.83	0.82	0.84	0.85	0.78	**0.83**
EFFNet [14]	0.84	0.83	**0.86**	0.86	0.80	**0.83**
4D-MENet(ours)	**0.85**	**0.85**	0.84	**0.90**	0.80	0.81

[1] Inflated 3D ConvNet. [2] No data augmentation. [3] Frame perturbation. [4] Temporal shift module.[5] Long short-term memory layer. [6] Non-local module. [7] Timeception layer.

Table 3. Results of different methods on RISE.

Different Methods	Params	FLOPs	Latency	FPS
I3D [18]	12.3 M	62.7 G	30.56 ms	32.71
I3D-TSM [31]	12.3 M	62.7 G	31.85 ms	31.40
I3D-LSTM [18]	38.0 M	62.9 G	31.01 ms	32.25
I3D-NL [32]	12.3 M	62.7 G	30.32 ms	32.98
I3D-TC [33]	12.3 M	62.7 G	30.41 ms	32.88
EFFNet [14]	27.2 M	34.6 G	23.49 ms	42.57
4D-MENet (ours)	32.1 M	62.9 G	30.71 ms	32.88

3.3.2. Qualitative Analysis

To better assess the performance of this paper's algorithm, this paper provides a qualitative analysis of 11 current advanced algorithms which can intuitively reflect the specific characteristics of the algorithmic classification and understand the nature of this paper's algorithmic improvement.

Figure 4 shows a heat map comparing the different algorithms, where the first to fifth rows show the test results of the different algorithms on the Forest Smoke dataset, and the sixth to seventh rows show the test results of the different algorithms on the RISE dataset. The first image features small smoke targets with fog and cloud interference. The second image features light smoke targets with haze and cloud interference. The third image features small smoke targets with cloud and forest interference. The fourth image features small smoke targets with lighting and road interference. The fifth image features light smoke targets with rooftop and cloud interference. The sixth image features small smoke targets with shade and chimney interference. The seventh image features small smoke targets with water vapor interference. From these experimental results, it is evident that the proposed algorithm outperforms the second-best algorithm, EFFNet, by effectively eliminating interference from similar objects and more accurately identifying smoke regions.

Based on the above analysis, the 4D-MENet network proposed in this paper performs better for early smoke recognition in natural forest fire scenarios. It solves the problems of low recall and high false alarm rates that exist at present. This is mainly due to the following factors: (1) the important frame ordering module proposed in this paper eliminates the operation of the neural network to extract the same features repeatedly, which makes the neural network extract the smoke motion features well, thus, increasing the accuracy of smoke recognition. (2) The 4D-ME module proposed in this paper makes the neural network pay more attention to the motion region in the video, thus, distinguishing the interference

of the objects that are similar to the appearance of the smoke well and improving the feature extraction capability of light smoke targets, which increases the accuracy of smoke recognition and reduces the false alarm rate. (3) The high-resolution multi-scale fusion module proposed in this paper increases the network image recognition layer for small targets, increasing smoke recognition accuracy.

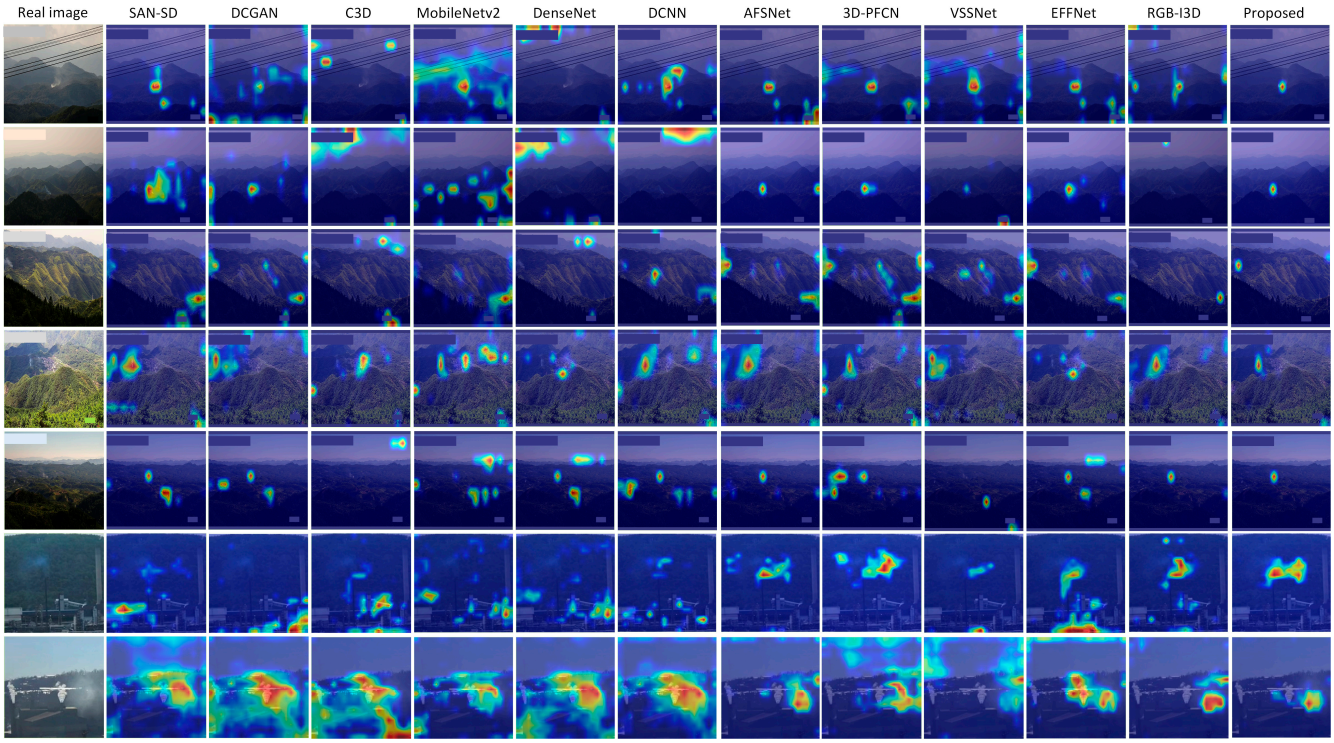

Figure 4. Heat map of different algorithms for early smoke recognition in real forest fire scenarios. In the figure, different colors represent the probability of the target's presence. Red indicates areas with a high likelihood, orange and yellow represent areas with a medium probability, green indicates areas with a low probability, and blue and purple represent areas with a very low probability.

3.4. Ablation Experiments

The advantages of the 4D-MENet network proposed in this paper are mainly the FS module, the 4D-ME module, and the HFM module. To validate the contributions of the different modules, the following ablation experiments are conducted in this section for comparative analysis.

3.4.1. Important Frame Sorting Module

To eliminate the negative impact of the network repeatedly extracting the same features, this paper proposes an FS. Section 2.3.2 provides a detailed description of the network architecture of this module. To validate its effectiveness, ablation experiments were conducted on the FS module, with specific experimental details and fine-tuning parameter settings provided in Section 3.1. The experiments were performed on 4D-MENet with and without the FS module on the Forest Smoke dataset. The network performance was evaluated using R, FPR, and F1-score metrics. Additionally, the FS module was fine-tuned with different frame sequence O parameters, and the network performance of the 4D-MENet on the Forest Smoke dataset under different O parameters was compared, as shown in Table 4.

From the analysis of the above table, it can be seen that introducing the FS significantly increases the recall rate, with only a slight increase in the false alarm rate. As the number of important frames continues to decrease, the network's performance continues to improve. It reaches its peak at **O** = 6, but as it continues to shrink, the performance metrics begin to decline, which may be caused by the extreme loss of the time dimension.

Table 4. Comparison results of our model variants with and without FS.

FS	O	R (%)	FAR (%)	F1-Score
no		93.75	5.12	0.9312
yes	12	94.24	4.78	0.9357
	10	94.68	4.42	0.9368
	8	94.92	4.38	0.9412
	6	**95.62**	**4.12**	**0.9486**
	4	94.54	4.26	0.9433

3.4.2. 4D Attention-Based Motion Target Enhancement Module

To distinguish the interference of objects similar to the appearance of smoke and improve the representation of light smoke features, this paper proposes the 4D-ME module. In Section 2.3.3, the network architecture of this module is described in detail. To validate its effectiveness, ablation experiments were conducted on the 4D-ME module, with specific experimental details and fine-tuning parameter settings provided in Section 3.1. Experiments were performed on 4D-MENet with and without the 4D-ME module on the Forest Smoke dataset. The network performance was evaluated using R, FPR, and F1-score metrics. Additionally, different attention mechanism A parameters in the 4D-ME module were fine-tuned, and the network performance of 4D-MENet on the Forest Smoke dataset under different A parameters was compared, as shown in Table 5.

Table 5. Comparison results of our model variants with and without 4D-ME.

4D-ME	A	R (%)	FAR (%)	F1-Score
no		94.25	7.12	0.9309
yes	S_a	94.86	5.68	0.9366
	C_a	94.54	6.22	0.9338
	T_a	95.00	5.04	0.9428
	$S_a \times C_a$	95.16	5.24	0.9404
	$C_a \times T_a$	95.24	4.82	0.9412
	$T_a \times S_a$	95.48	4.54	0.9462
	$\mathbf{S_a \times C_a \times T_a}$	**95.62**	**4.12**	**0.9486**

The above table shows a slight increase in the recall rate. In contrast, the false alarm rate has a more substantial increase, indicating that the module significantly improves similar object differentiation and, to a certain extent, improves the feature extraction ability of light smoke targets. The comparison data of different attention mechanisms in Table 5 reveal that temporal attention mechanisms significantly impact the performance of smoke detection. This is primarily because the dynamic changes in smoke are a critical feature for smoke target detection. The network achieves optimal performance when spatial attention, channel attention, and temporal attention mechanisms are combined.

3.4.3. High-Resolution Multi-Scale Fusion Module

To increase the small target image recognition layer and improve the semantic information and spatial information of high-level feature maps used for classification, this paper proposes the HFM module. Section 2.3.4 provides a detailed description of the network architecture of the HFM module. To validate its effectiveness, ablation experiments were conducted on the HFM module, with specific experimental details and fine-tuning parameter settings provided in Section 3.1. Experiments were performed on 4D-MENet with and without the HFM module on the Forest Smoke dataset. The network performance was evaluated using the R, FPR, and F1-score metrics. Additionally, the parameter for the number of repetitions, N, in the HFM module was fine-tuned, and the network performance of 4D-MENet on the Forest Smoke dataset with different N parameters was compared, as shown in Table 6.

Table 6. Comparison results of our model variants with and without HMF.

HFM	N	R (%)	FAR (%)	F1-Score
no		94.52	4.78	0.9324
	3	94.63	4.56	0.9359
yes	**4**	**95.62**	**4.12**	**0.9486**
	5	95.54	4.28	0.9412

From the above table, it can be seen that the introduction of HMF significantly increases the recall rate. As the number of repetitions of the HMF module increases, the network's performance continues to improve and reaches its peak when N = 4. However, the performance metrics start to decrease when N continues to grow, which may be due to the overfitting of the network.

3.5. Practical Application

To validate the practical application of the proposed algorithm in real-world scenarios, this study selected a small number of smoke samples from diverse indoor and outdoor scenes in publicly available datasets, including CVPR [34], USTC [35], XJTU-RS [36], and Kaggle wildfire smoke [37]. Due to the absence of consistent labels across these smoke samples, this paper directly compares the visual results of different smoke instances. During experimentation, the model trained on a custom dataset is utilized, and consistent testing parameters are applied to evaluate the samples from each dataset. Detailed experimental procedures are elaborated on in Section 3.1. The specific experimental results are illustrated in Figure 5. Figure 5 shows that the proposed algorithm exhibits good network performance in different scenarios, including indoor cigarette smoke, outdoor smoke, and synthesized smoke. Therefore, this algorithm plays a crucial role in indoor and outdoor fire prevention.

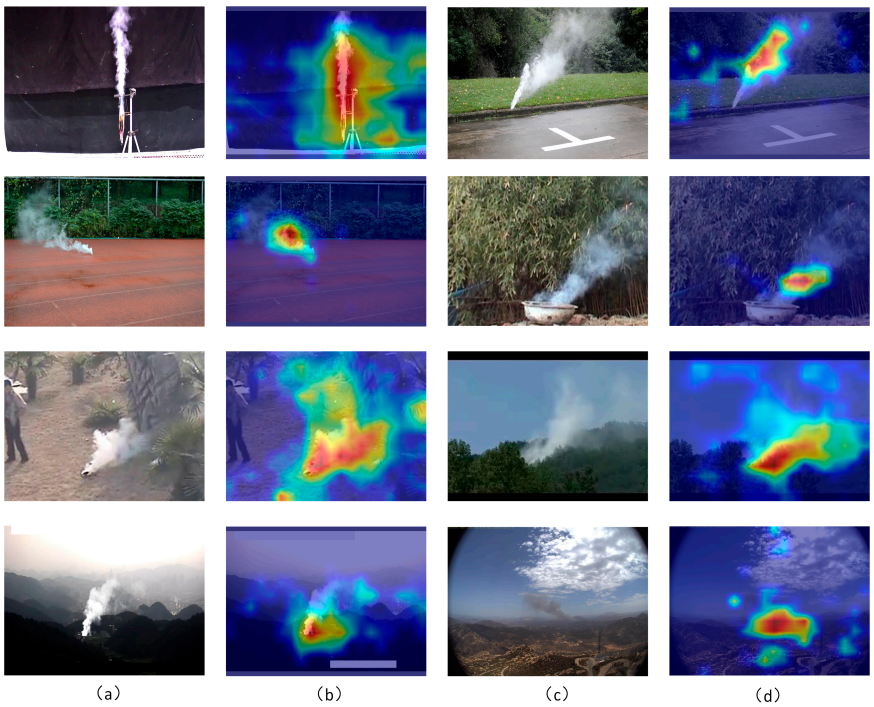

Figure 5. Visualization of indoor and outdoor smoke target detection results in different scenarios. Columns (**a**,**c**) show smoke samples from other scenes and (**b**,**d**) show corresponding heat maps. In the figure, different colors represent the probability of the target's presence. Red indicates areas with a high likelihood, orange and yellow represent areas with a medium probability, green indicates areas with a low probability, and blue and purple represent areas with a very low probability.

4. Discussion

Forests cover one-third of the Earth's land area, and trees absorb a significant amount of carbon dioxide through photosynthesis, playing a crucial role in maintaining the health of the Earth's environment. However, forest fires have devastating impacts on local residents' lives and on terrestrial environments, causing irreparable damage to atmospheric conditions and ecosystems. Therefore, timely prevention of forest fires is vital for both ecological environments and human societies. Forest fires typically generate smoke initially, making smoke detection crucial for early fire detection and for reducing fire losses. However, current forest fire smoke detection systems suffer from low recall rates and high false alarm rates. A low recall rate means many actual fires go undetected, rendering fire prevention efforts almost meaningless, while a high false alarm rate increases the workload of firefighting personnel. Hence, further research and solutions are needed to improve the reliability and efficiency of detection systems.

Currently, mainstream smoke detection algorithms include SAN-SD, EFFNet, VSSNet, 3D-PFCN, I3D, CNN-LSTM, DCNN, MobileNetv2, DenseNet, DCGAN, C3D, MOG-CNN, and others. In this study, experiments were conducted on different smoke detection algorithms using both self-made datasets and public datasets. The experimental results are shown in Tables 1–3, as well as in Figure 4. From qualitative and quantitative analysis, it can be seen that among all mainstream algorithms, the proposed algorithm exhibits the best network performance, with EFFNet being the second-best algorithm. The main reasons for the poorer performance of other algorithms are as follows:

Reason 1: Although all of the above algorithms have ideal recognition results on their datasets, the recognition results will be abruptly downgraded if they are changed to a different dataset, mainly because there is no available large-scale smoke dataset, which leads to poor performance in detecting scenarios that do not exist in the dataset.

Reason 2: Existing algorithms cannot recognize small smoke targets. The main reason is that during the feature extraction process of the neural network, as the depth increases, shallow features (small target features) will gradually be covered by high-level features, and high-level features have strong semantics. Small smoke target information will be lost when performing category recognition on high-level features.

Reason 3: Existing algorithms cannot recognize slow-moving smoke targets. The main reason is that the slow movement of smoke will cause several consecutive frames in the video to have the same features. When the neural network repeatedly extracts the same features, it will harm network prediction and is not conducive to model training and convergence.

Reason 4: Existing algorithms cannot recognize light smoke targets. The main reason is that the color becomes translucent as the smoke rises, making it easy for the neural network to ignore this part of the content when extracting features, resulting in the extracted features not having discriminative properties.

Reason 5: Existing algorithms cannot recognize smoke targets in complex backgrounds. The main reason is that there are many interferences between the appearance of smoke and familiar targets in the forest background (clouds, fog, haze, roads, roofs, etc.), resulting in the network being unable to extract discriminative features.

Based on the analysis above, this paper proposes the 4D-MENet network, which incorporates three additional modules, FS, 4D-ME, and HFM, into the RGB-I3D algorithm, significantly enhancing network performance. Compared to EFFNet, this paper first replaces the random keyframe extraction method with an adaptive keyframe extraction method, reducing redundancy and highlighting important information. Subsequently, the 2D attention mechanism in the intermediate feature layer of EFFNet is replaced with a 4D attention mechanism to enhance the smoke's temporal characteristics and channel feature representation capabilities. Finally, a small target detection layer is added, resulting in the improved network performance of the proposed algorithm. Tables 1 and 2 show that compared to EFFNet, the proposed algorithm exhibits significant improvements in both recall rate and false alarm rate network performance. Visualization in Figure 4 demonstrates

that the proposed algorithm is superior from the subjective perspective. Furthermore, in Section 3.5, this paper conducts experiments using the proposed algorithm on various publicly available datasets (including indoor, outdoor, and synthesized smoke) in the same experimental environment. The experimental results are visualized to demonstrate that the proposed algorithm can be applied to smoke detection in forest scenes and directly to indoor and outdoor smoke detection in non-forest scenes. Therefore, the proposed algorithm is significant in terms of applicability across various scenarios. However, the algorithm proposed in this paper has certain limitations regarding real-time performance, particularly for tasks such as real-time monitoring of smoke in forest fire videos, where high real-time requirements are crucial. A dependable smoke video detection algorithm must meet specific performance metrics: FLOPs between 10 G and 100 G, FPS above 30 fps, and model size ranging from 5 M to 50 M. As shown in Table 3, the proposed algorithm performs reliably on these metrics. However, compared to EFFNet, our algorithm lags by 28.3 G FLOPs, 9.69 fps in FPS, and 4.9 M in model size. Subsequent optimization of the proposed algorithms is necessary. In resource-constrained environments, the following optimization strategies are recommended:

Parallel processing: Use multi-threading or GPU acceleration to process different frames or image regions in parallel, reducing the processing time per frame.

Algorithm simplification: Employ more straightforward and more efficient algorithmic steps, such as replacing morphological processing with lighter-weight filtering techniques.

Frame rate adjustment: In extremely resource-constrained scenarios, reducing the video frame rate (e.g., from 30 FPS to 15 FPS) can decrease the number of frames processed per second, thus, alleviating the computational burden.

Region of interest (ROI): Process only the regions of interest (ROIs) in the video, ignoring static areas, thereby reducing the processing load.

Model compression: Reduce computational and storage requirements through model compression techniques such as pruning and quantization.

5. Conclusions

To solve the early smoke recognition difficulties that currently exist with forest fires, this paper proposes 4D-MENet, which introduces three modules: FS, 4D-ME, and HFM on the RGB-I3D network, significantly improving network performance. Meanwhile, this paper integrates a larger-scale video dataset of early smoke from forest fires in real scenarios, labeled with 6250 video classification labels, laying a better foundation for future practical applications. Experimental results show the algorithm's effectiveness, achieving a 95.62% recall on the Forest Smoke dataset, a 4.12% false alarm rate, and outstanding F1-scores on the three subsets of the RISE dataset (0.85, 0.85, and 0.90, respectively). Looking ahead, future research will explore diverse algorithm implementations, including nighttime smoke recognition in challenging environmental scenarios. The smoke recognition algorithm introduced in this study demonstrates applicability beyond forest fire safety measures, extending its use to various scenarios, including urban and indoor fire incidents. This broader application scope aims to mitigate the impact of fires on public infrastructure and human well-being.

Author Contributions: Conceptualization, Y.W. and Y.P.; methodology, Y.W.; software, Y.W. and H.W.; validation, Y.W. and H.Z.; formal analysis, Y.W.; writing—original draft, Y.W.; writing—review and editing, Y.W.; visualization, Y.W.; supervision, Y.P.; resources, Y.P.; project administration, Y.P.; funding acquisition, Y.P.; data curation, Q.W. and N.Q.; investigation, Q.W. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Project of the Jilin Provincial Department of Science and Technology (YDZJ202402041CXJD; 20240101359JC).

Data Availability Statement: The data introduced in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gürbüz, H.; Şöhret, Y.; Ekici, S. Evaluating effects of the Covid-19 pandemic period on energy consumption and enviro-economic indicators of Turkish road transportation. *Energy Sources Part A Recovery Util. Environ. Eff.* **2021**, 1–13. [CrossRef]
2. Ekici, S.; Şöhret, Y.; Gürbüz, H. Influence of COVID-19 on air pollution caused by commercial flights in Turkey. *Energy Sources Part A Recovery Util. Environ. Eff.* **2021**, 1–13. [CrossRef]
3. Barmpoutis, P.; Dimitropoulos, K.; Grammalidis, N. Smoke detection using spatio-temporal analysis, motion modeling and dynamic texture recognition. In Proceedings of the 2014 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 13 November 2014; pp. 1078–1082.
4. Dimitropoulos, K.; Barmpoutis, P.; Grammalidis, N. Higher order linear dynamical systems for smoke detection in video surveillance applications. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1143–1154. [CrossRef]
5. Islam, M.R.; Amiruzzaman, M.; Nasim, S.; Shin, J. Smoke object segmentation and the dynamic growth feature model for video-based smoke detection systems. *Symmetry* **2020**, *12*, 1075. [CrossRef]
6. Wu, X.; Cao, Y.; Lu, X.; Leung, H. Patchwise dictionary learning for video forest fire smoke detection in wavelet domain. *Neural Comput. Appl.* **2021**, *33*, 7965–7977. [CrossRef]
7. Hu, Y.; Lu, X. Real-time video fire smoke recognition by utilizing spatialtemporal convnet features. *Multimed. Tools Appl.* **2018**, *77*, 29283–29301. [CrossRef]
8. Aslan, S.; Gudukbay, U.; Toreyin, B.U.; Etin, A.E. Early wild-fire smoke recognition based on motion-based geometric image transformation and deep convolutional generative adversarial networks. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 17 April 2019; pp. 8315–8319.
9. Yang, X.; Sun, Y. Research on smoke recognition based on densenet. In Proceedings of the 2019 ACM Southeast Conference, Kennesaw, GA, USA, 18–20 April 2019; pp. 160–163.
10. Hsu, Y.C.; Hao, T.; Huang; Hu, T.Y.; Nourbakhsh, I. RISE video dataset: Recognizing industrial smoke emissions. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually. 2–9 February 2021; pp. 14813–14821.
11. Shi, J.; Wang, W.; Gao, Y.; Yu, N. Optimal placement and intelligent smoke recognition algorithm for wildfire-monitoring cameras. *IEEE Access* **2020**, *8*, 72326–72339. [CrossRef]
12. Tao, H.; Duan, Q. An adaptive frame selection network with enhanced dilated convolution for video smoke recognition. *Expert Syst. Appl.* **2023**, *215*, 119371. [CrossRef]
13. Jiang, M.; Zhao, Y.; Yu, F.; Zhou, C.; Peng, T. A self-attention network for smoke recognition. *Fire Saf. J.* **2022**, *129*, 103547. [CrossRef]
14. Cao, Y.; Tang, Q.; Wu, X.; Lu, X. EFFNet: Enhanced Feature Foreground Network for Video Smoke Source Prediction and Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1820–1833. [CrossRef]
15. Li, X.; Chen, Z.; Wu, Q.M.J.; Liu, C. 3D Parallel fully convolutional networks for real- time video wildfire smoke recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 89–103. [CrossRef]
16. Zhu, G.; Chen, Z.; Liu, C.; Rong, X.; He, W. 3D video semantic segmentation for wildfire smoke. *Mach. Vis. Appl.* **2020**, *31*, 50. [CrossRef]
17. Quan, Y.; Zhang, D.; Zhang, L.; Tang, J. Centralized Feature Pyramid or Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 4341–4354. [CrossRef] [PubMed]
18. Carreira, J.; Zisserman, A. Quo Vadis. Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
19. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
21. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1063–6919.
23. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
25. Ghiasi, G.; Lin, T.Y.; Pang, R.; Le, Q.V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038.
26. Tan, M.; Le, Q.V. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
27. Cao, Y.; Lu, X. Learning spatial-temporal representation for smoke vehicle Detection. *Multimed. Tools Appl.* **2019**, *78*, 27871–27889. [CrossRef]

28. Gu, K.; Xia, Z.; Qiao, J.; Lin, W. Deep dual-channel neural network for image-based smoke recognition. *IEEE Trans. Multimed.* **2019**, *22*, 311–323. [CrossRef]
29. Lin, G.; Zhang, Y.; Xu, G.; Zhang, Q. Smoke recognition on video sequences using 3D convolutional neural networks. *Fire Technol.* **2019**, *55*, 1827–1847. [CrossRef]
30. Nguyen, M.D.; Kim, D.; Ro, S. A video smoke recognition algorithm based on cascade classification and deep learning. *KSII Trans. Internet Inf. Syst. (TIIS)* **2018**, *12*, 6018–6033.
31. Lin, J.; Gan, C.; Han, S. TSM: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7083–7093.
32. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
33. Hussein, N.; Gavves, E.; Smeulders, A. Timeception for complex action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 254–263.
34. Ko, B.; Ham, S.; Nam, J. Modeling and formalization of fuzzy finite automata for detection of irregular fire flames. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1903–1912. [CrossRef]
35. Zhang, Q.; Lin, G.; Zhang, Y.; Xu, G.; Wang, J. Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. *Procedia Eng.* **2018**, *211*, 441–446. [CrossRef]
36. Wang, J.; Zhang, X.; Jing, K.; Zhang, C. Learning precise feature via self-attention and self-cooperation YOLOX for smoke detection. *Expert Syst. Appl.* **2023**, *228*, 120330. [CrossRef]
37. Al-Smadi, Y.; Alauthman, M.; Al-Qerem, A.; Aldweesh, A.; Quaddoura, R.; Aburub, F.; Mansour, K.; Alhmiedat, T. Early wildfire smoke detection using different yolo models. *Machines* **2023**, *11*, 246. [CrossRef]