IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 20, 2019, accepted October 4, 2019, date of publication October 11, 2019, date of current version November 5, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2946712*

# An Attention Enhanced Bidirectional LSTM for Early Forest Fire Smoke Recognition

**YICHAO CAO[1,2], FENG YANG[3], QINGFEI TANG[4], AND XIAOBO LU[1,2]**

[1]School of Automation, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China
[3]College of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[4]Nanjing Enbo Technology Company Ltd., Nanjing 210007, China

Corresponding author: Xiaobo Lu (xblu2013@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871123, in part by the Key Research and Development Program in Jiangsu Province BE2016739, and in part by the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

**ABSTRACT** Detecting forest fire smoke during the initial stages is vital for preventing forest fire events. Recent studies have shown that exploring spatial and temporal features of the image sequence is important for this task. Nevertheless, since the long distance wildfire smoke usually move slowly and lacks salient features, accurate smoke detection is still a challenging task. In this paper, we propose a novel Attention Enhanced Bidirectional Long Short-Term Memory Network (ABi-LSTM) for video based forest fire smoke recognition. The proposed ABi-LSTM consists of the spatial features extraction network, the Bidirectional Long Short-Term Memory Network(LSTM), and the temporal attention subnetwork, which can not only capture discriminative spatiotemporal features from image patch sequences but also pay different levels of attention to different patches. Experiments show that out ABi-LSTM is capable of achieving best accuracy and less false alarms on different types of scenarios. The ABi-LSTM model achieve a highly accuracy of 97.8%, and there is 4.4% improvement over the image-based deep learning model.

**INDEX TERMS** Smoke detection, attention, LSTM, spatiotemporal features.

## I. INTRODUCTION

An efficient and stable vision-based smoke detection algorithm is critical for the initial forest fire detection. On one hand, forest fires present a significant challenge to human life and natural ecological environment. If a forest fire cannot be promptly extinguished, it will have a bad impact on a wide area. Reaction time is one of the key factors that determine the success of forest fire suppression. On the other hand, there were extensive research on photoelectric- or ionization-based fire smoke detectors. However, these sensors are limited by the fact that these always serve as point sensors in space, which are unsuitable at monitoring larger areas such as early forest fire detection. The limitations of current smoke sensors have prompted researches on vision-based smoke detection methods.

Pan-tilt-zoom (PTZ) IP cameras are excellent for viewing large areas. They can be placed in auto-patrol modes where they automatically step through predetermined positions. This paper proposes a novel methods to detect forest fire using PTZ IP cameras. Figure 1 illustrates the pan-tilt-zoom (PTZ) long range camera for forest fire detection and a snapshot of a typical forest fire smoke at the initial stages captured by a forest watch tower. The main manifestation of early forest fires is smoke because of tree shelter and terrain. Therefore, forest fire monitoring system always focus on smoke identification.

A considerable volume of research effort within the last decade focused mainly on the identification of specific features of smoke. Existing methods of smoke detection can be divided into two categories: image-based smoke detection [1]–[3] and video-based smoke detection [4], [5]. The general smoke detection algorithms usually combine motion detection, feature extraction and classification method. Image-based smoke detection methods are usually independent of inter-frame context information. Video-based methods usually not only analyze spatial features in single frame images, but also extract temporal features between frames.

Under certain conditions, single-frame-based detection method is a good choice when it is difficult to obtain stable

The associate editor coordinating the review of this manuscript and approving it for publication was Xian Sun.

**FIGURE 1.** A pan-tilt-zoom (PTZ) long range camera for forest fire detection and a snapshot of a typical forest fire smoke at the initial stage captured by a forest watch tower.

and reliable image sequence. Tian *et al.* [1] recently proposed to separate a frame into quasi-smoke and quasi-background components by convex optimization. Deep learning with convolutional neural networks (CNNs) has achieved great success in image classification and target detection. In [2], researchers proposed a deep normalization and convolutional neural network (DNCNN) with 14 layers to implement automatic feature extraction and classification. Yuan *et al.* [3] proposed a smoke detection method that combines local binary pattern (LBP) like features, kernel principal component analysis (KPCA), and Gaussian process regression (GPR).

However, dynamic feature is one of the essential features in forest fire smoke recognition task. The human vision system is incredibly good at recognizing complex moving smoke in sequence image, because it analyzes dynamic characteristics when judging. If dynamic features can be extracted and modeled better, it would be helpful for improving the recognition accuracy. Dimitropoulos *et al.* [4] introduced a higher order linear dynamical system (h-LDS) descriptor for multidimensional dynamic texture analysis. There are also researchers applying deep learning to forest fireworks identification. Lin *et al.* [5] proposed a joint detection framework based on faster RCNN and 3D CNN. However, the application of this algorithm is restricted by the large computational complexity in practice.

Because the moving speed and direction of smoke in the image are related to the monitoring distance and weather, it is necessary for the model to adapt to a variety of scenes. The difficulties of accurate forest fire smoke recognition lie in two aspects, (1) learning efficient spatiotemporal representation of fire smoke; (2) early forest fire smoke has different motion saliency in different frames, so the model should pay different attention to each frame.

Given the aforementioned concerns, we propose our novel attention enhanced bidirectional LSTM Network (ABi-LSTM) for forest fire smoke recognition. The foreground detection algorithm is used to extract candidate image patch sequences from video. And the block-based detection scheme is used to expand the recognition scope (the background information around the motion pixels can be obtained effectively) and roughly locate the smoke fire area.

This paper focuses on the candidate image patch sequences classification. The contributions of this paper are summarized as follows:

- We propose a novel attention enhanced bidirectional LSTM network (ABi-LSTM) to tackle the early forest fire smoke recognition problem.
- We consider spatiotemporal representation of smoke candidate patch by applying CNN and bidirectional long short-term memory network from forward and backward time direction.
- This is the first publication to apply attention mechanism for video-based forest fire smoke recognition. In our specific implementation, an attention network is designed to self-adaptively focus on discriminative frames with a soft attention mechanism that can automatically emphasize motion information in temporal domain.
- We construct more challenging forest fire smoke data sets to increase the reliability of the experiment. Experimental results demonstrate that the proposed method outperforms existing methods for forest fire smoke recognition.

The rest of this paper is organized as follows. The proposed ABi-LSTM framework is described in Section 3. The first part of this section describes the spatial features extraction network, which is actually an Inception V3 network [23]; the second part briefly review the Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) and build a multi-layer bidirectional LSTM model by feed spatial

feature of single patch to extract temporal features from forward and backward order; the third part proposes an attention network to optimize classification process with a soft attention mechanism. In Section 4, experimental results are presented, and the ABi-LSTM framework is compared with other smoke recognition algorithms. Finally, conclusions are drawn in Section 5.

## II. RELATED WORK

Although there is little literature on early forest fire smoke detection, there is substantial literature on video based smoke detection and fire detection [6], [7]. Many researchers have attempted to address the problem of smoke detection focusing mainly on the recognition of spatiotemporal features of smoke. As mentioned in the previous section, existing methods of smoke detection can be divided into two categories: image-based smoke method and video-based smoke method.

### A. IMAGE-BASED METHOD

From the point of view of image-based method, there is a vast literature about the investigations of the static characteristics of smoke. Inspired by the airlight-albedo ambiguity model, a novel approach to detect smoke using transmission is proposed in [8]. In order to improve the performance, Yuan [9] proposed a double mapping framework which concatenates histograms of edge orientation, edge magnitude and Local Binary Pattern (LBP) bit, and densities of edge magnitude, LBP bit, color intensity and saturation. Tian *et al.* [1], [10], [11] formulated the smoke separation problem as convex optimization that solves a sparse representation problem. In [10], three different models that constrain the smoke component are proposed to separate the smoke component from a given frame. In [11], the sparse coefficients associated with an over-complete dictionary representation is used to detect smoke as a new feature. Furthermore, Tian *et al.* [1] solved the sparse representation problem using dual dictionaries for the smoke and background components, respectively, and developed a method based on the concept of image matting to separate the smoke and background components from a single image frame.

However, important dynamic information is often lost in a single frame image, which is one of the main reasons for the difficulty of image-based method.

### B. VIDEO-BASED METHOD

As one of essential features, the motion information of smoke will undoubtedly improve the smoke recognition accuracy in theory. In [12], a smoke detection method using color, motion and growth properties are proposed. Dimitropoulos *et al.* [4] introduced a higher order linear dynamical system (h-LDS) descriptor to analyze the smoke candidate image patches in each subsequence. Undoubtedly, the extraction of dynamic information in the process of recognition improves the performance of mode to some extent.

The above methods usually focus on the boundary of smoke or the effects of smoke on the edges of objects covered by smoke by hand crafted features. Traditional hand crafted feature based smoke detection methods can achieve high accuracy in a small amount of samples but generalization performances are less than satisfactory due to sensitivity to the parameter setting of the detection algorithm. Moreover, hand crafted feature based methods usually recognize smoke from small size blocks (often $< 50 \times 50$), which limits the accuracy of smoke recognition.

### C. DEEP LEARNING METHOD

In recent years, Deep Learning approaches (e.g. Convolutional Neural Networks and Recurrent Neural Networks) has led to very good performance on a variety of problems, such as visual recognition [13], speech recognition [14] and natural language processing [15]. Yin *et al.* [2] proposed a deep normalization and convolutional neural network (DNCNN) with batch normalization to extract features for smoke detection. In [16], researchers demonstrated the effectiveness of saliency detection method and CNN in localization and recognition of wildfire in aerial images. Liu *et al.* [17] proposed a dual convolution network using dark channel prior (DarkC-DCN) to further improve the recognition accuracy of image-based CNN model. To ease the limitations of smoke image samples, an end-to-end trainable framework based on fast detector SSD and MSCNN for smoke detection is proposed, which can optimize the model from synthetic and real smoke samples.

Moreover, there is also video-based method using deep learning [5]. A joint detection framework based on faster RCNN [19] and 3D CNN [20] is proposed to detection smoke, in which an improved faster RCNN with non-maximum annexation is responsible for the smoke target location and 3D CNN is responsible for smoke recognition by combining dynamic spatial–temporal information. Although this video-based method takes into account the dynamic characteristics between different frames, it can hardly be used in practical scenarios because of the high computational cost.

Besides CNN, Recurrent Neural Network (RNN) is another important structure of deep learning, which has made significant breakthroughs in various tasks, especially sequence processing [21]. However, the vanishing gradient problem is a difficulty found in training recurrent neural network with Back-Propagation Through Time. Long Short Term Memory (LSTM) is specifically designed to tackle this problems [22], [24]. There have been some meaningful works about RNN and LSTM [32]–[34]. Attention mechanism is another most influential ideas in the Deep Learning community, which is used in various problems like neural machine translation, human action recognition and so on [25], [26]. The attention mechanism can focus on discriminative features in a longer sequence, which can be used in many difficult tasks.

Our key motivation of ABi-LSTM is that: a) compared with the hand crafted feature, CNN has more powerful feature extraction ability, and the block size used for forest fire smoke recognition in this paper is larger, which is helpful for CNN

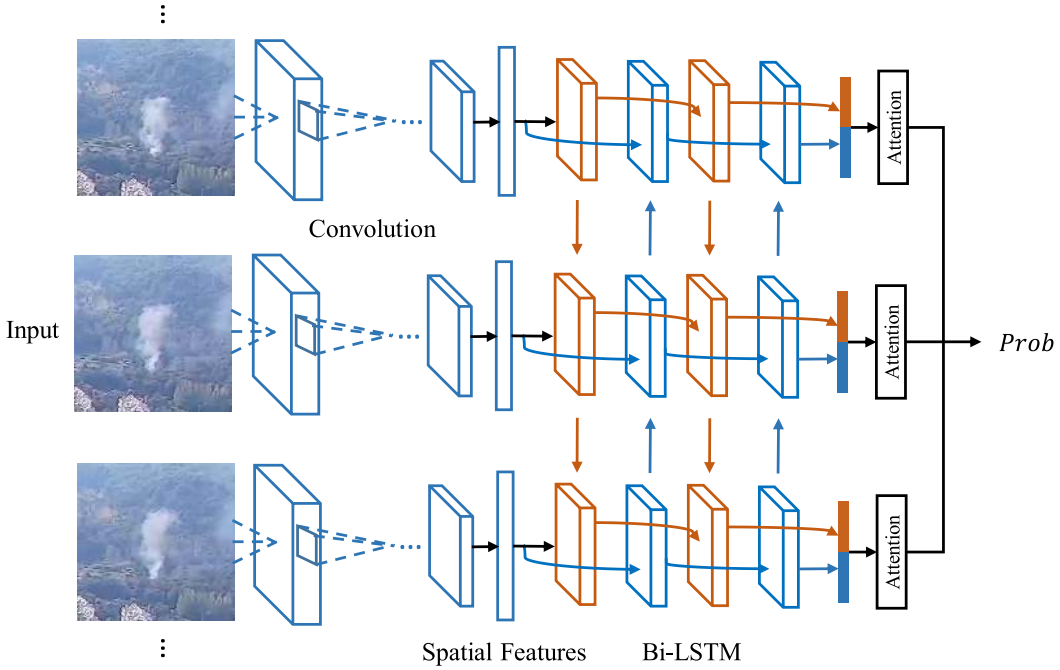


**FIGURE 2.** Framework of the proposed ABi-LSTM for forest fire smoke recognition, which consists of the spatial features extraction network, the bidirectional LSTM network, and the temporal attention subnetwork. The input images are fed into the ABi-LSTM network one by one.

model to refer to the surrounding information of suspected fire smoke in the prediction process; b) bidirectional LSTM can learn captures long-term information from forward and backward time direction; c) attention mechanism can guide the classification network focus on key frames from a long image sequences.

## III. APPROACH

In this section, we propose a novel Attention Enhanced Bidirectional LSTM architecture for forest fire smoke recognition.

### A. OVERVIEW OF METHODOLOGY

As illustrated in Figure 2, the proposed ABi-LSTM is mainly composed of three components: the spatial features extraction network, the Bidirectional LSTM network, and the temporal attention subnetwork. The spatial features extraction network is employed to extract spatial features from candidate patches, which are captured by ViBe [32] background subtraction method. The Bidirectional LSTM network learns long-term smoke-related information from spatial features. In order to make full use of both the past and future context information of a sequence in classification, a bidirectional LSTM is employed to extract temporal features from forward and backward order. In this model, the orange arrows indicate the direction of information flow in forward LSTM and the blue arrows indicate the direction of information flow in backward LSTM. In order to concentrate on discriminative frames which contribute more on forest fire smoke recognition, an attention subnetwork is designed to automatically emphasize motion information with a soft attention mechanism in temporal domain. We'll provide a detailed explanation of each component later.

### B. SPATIAL FEATURES EXTRACTION

CNNs have achieved excellent performance in computer vision tasks. The Inception network was an important milestone in the development of CNN classifiers. GoogLenet is known as Inception V1 [27], and the researchers have subsequently proposed improved models such as Inception V2 [28] and Inception V3 [23]. In this paper, instead of building a model from scratch, a pretrained Inception V3 model is used to capture spatial information from each individual frame.

Inception V3 is a heavily engineered network, which used a lot of upgrades to increase the accuracy and reduce the computational complexity: (1) Factorize $5 \times 5$ convolution to two 3x3 convolution operations to improve computational speed. (2) Factorize $n \times n$ convolution to a combination of $1 \times n$ and $n \times 1$ convolutions. (3) Expand the filter bank outputs to remove the representational bottleneck. (4) Combination of additional regularization with batch-normalized auxiliary classifiers and label-smoothing.

In this study, the output of the ''avg_pool'' layer of Inception V3 is used as spatial feature instead of the fully-connected layer. The 2048-dimensional image features at each time-step will form spatial features sequence that are learned by subsequent bidirectional LSTM.

### C. BIDIRECTIONAL LSTM

In this section, we briefly review the Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) to make the paper self-contained. RNN is an extension of

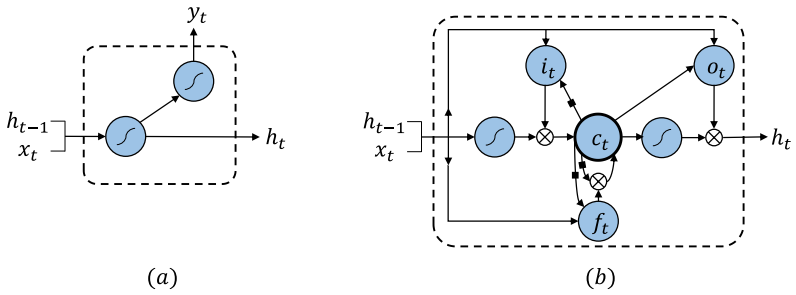


**FIGURE 3.** Structures of the neurons. (a) RNN and (b) LSTM.

feed-forward neural networks and has yielded promising results in sequence learning. Figure 3-a demonstrates an RNN neuron. The input of the RNN is a sequence data $\{x_1, x_2, \ldots, x_T\}$.

As shown in Figure 3-a, the hidden state of all RNN units at the $t$th time step is determined by the current input $X_t$ and the previous hidden state $h_{t-1}$ at the $(t-1)$th time step.

$$h_t = \sigma(W_{xh} \cdot X_t + W_{hh} \cdot h_{t-1} + b_h) \quad (1)$$

$$y_t = g(\sigma\,(W_{ho} \cdot h_t + b_o)) \quad (2)$$

where $\sigma$ is a nonlinear activation function, $g$ denotes the operation of the fully-connected layer, $b_h$ and $b_o$ are bias vectors, $W_{xh}$, $W_{hh}$ and $W_{ho}$ denote weight matrices from the current input layer to hidden layer, the previous hidden layer to current hidden layer and the current hidden layer to output layer, respectively. RNN is an important model for sequential date modeling of the deep learning family. However, it comes with some challenges in modelling long-term dependencies such as vanishing and exploding gradient problems during the training phase. Our model builds on LSTM cells, which is an advanced RNN architecture explicitly designed for tackling this problem. Our key motivation of chosen LSTM is that it can learn long-term dependencies and avoid exploding and vanishing gradient problems that traditional RNN suffers from during back propagation optimization. LSTM has been successfully applied to handwriting recognition, machine translation and so on. The difference between LSTM and RNN is that the later adds several gates to the cell to judge whether the information is useful or not [39].As illustrated in Figure 3-b, a LSTM neuron updates its memory cell state $C_t$ from different sources at given time step $t$: the current input $X_t$, the hidden state from LSTM themselves at the last time step $h_{t-1}$ as well as previous memory cell state $C_{t-1}$.

At each time step, the LSTM neuron can choose to input, forget, and output the memory cell state governed by four important parts: input gate $i_t$, output gate $o_t$,forget gate $f_t$ and candidate cell state $\tilde{C}_t$. Based on these parts, LSTM neuron memory cell state and output can be computed by:

$$i_t = \sigma(W_{xi} \cdot X_t + W_{hi} \cdot h_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf} \cdot X_t + W_{hf} \cdot h_{t-1} + b_f) \quad (4)$$

$$o_t = \sigma(W_{xo} \cdot X_t + W_{ho} \cdot h_{t-1} + b_o) \quad (5)$$

$$\tilde{C}_t = tanh(W_{hC} \cdot h_{t-1} + W_{xC} \cdot X_t + b_C) \quad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (7)$$

$$h_t = o_t \odot tanh(C_t) \quad (8)$$

where $tanh$ presents hyperbolic tangent function, '$\cdot$' is a matrix multiplication operator, '$\odot$' denotes the products with a gate value, and $b_i$, $b_f$, $b_o$ and $b_C$ are bias vectors. The weight matrix subscripts have obvious meaning. For example, $W_{hi}$, $W_{xo}$ and $W_{ho}$ denote hidden-input gate matrix, input-output gate matrix and hidden-output gate matrix, respectively. In the proposed ABi-LSTM, multi layers LSTM are stacked to learn long term dependencies in sequence data.

In order to make full use of both the past and future context information of a sequence in classification, we build a bidirectional LSTM model by feed spatial feature of single patch to extract temporal features from forward and backward order. The bidirectional LSTM model consists of two parts: forward LSTM and backward LSTM as illustrated in Figure 4. The forward LSTM updates its memory cell state $\vec{C}_t$, starting at time $t = 1$ (from $x_1$ to $x_T$). Similarly, the backward LSTM updates its memory cell state $\overleftarrow{C}_t$, starting at time $t = T$ (from $x_T$ to $x_1$). Formally, the bidirectional LSTM model works as follows, for raw image patch $I_t$, forward memory cell state $\vec{C}_t$ and backward memory cell state $\overleftarrow{C}_t$, the encoding performs as

$$X_t = \mathcal{C}(I_t, \Theta_{\mathcal{C}}), \quad \vec{C}_t = \vec{\mathcal{T}}(X_t, \Theta_{\vec{\mathcal{T}}}), \quad (9)$$

$$\overleftarrow{C}_t = \overleftarrow{\mathcal{T}}(X_t, \Theta_{\overleftarrow{\mathcal{T}}})$$

$$O_t = \mathcal{M}(X_t, \Theta_{\mathcal{M}}) \quad (10)$$

where $\mathcal{C}$, $\vec{\mathcal{T}}$, $\overleftarrow{\mathcal{T}}$ represent CNN, forward LSTM and backward LSTM respectively and $\Theta_{\mathcal{C}}$, $\Theta_{\vec{\mathcal{T}}}$ and $\Theta_{\overleftarrow{\mathcal{T}}}$ are their corresponding weights. $X_t$ is the spatial feature of a single frame extracted by CNN. $\mathcal{M}$ presents multi-layer LSTM and $\Theta_{\mathcal{M}}$ is multi-layer LSTM weights.

### D. ATTENTION MECHANISM

For a long image patch sequence, the amount of valuable information provided by different frames is in general not equal. We employ an attention network to adaptively focus

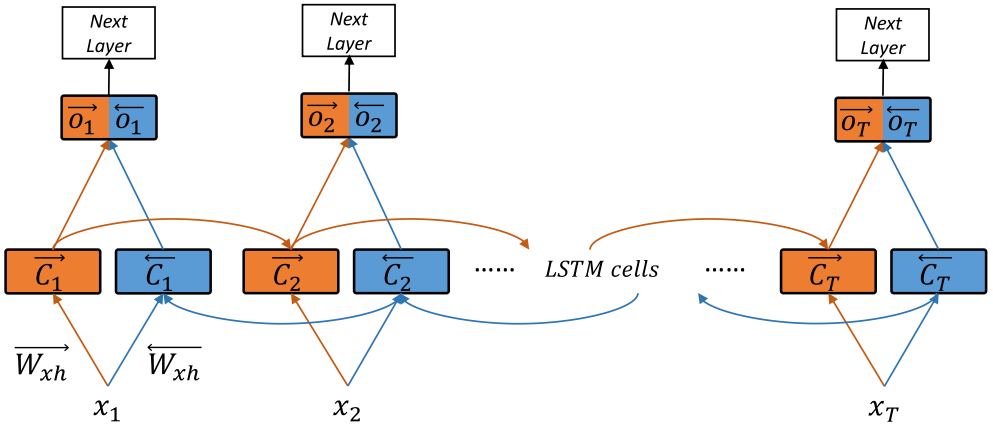


**FIGURE 4.** A single layer bidirectional LSTM. We feed spatial features in both forward (red arrows) and backward (blue arrows) order which allows our model learns both the past and future context information context information from both left and right side over time.

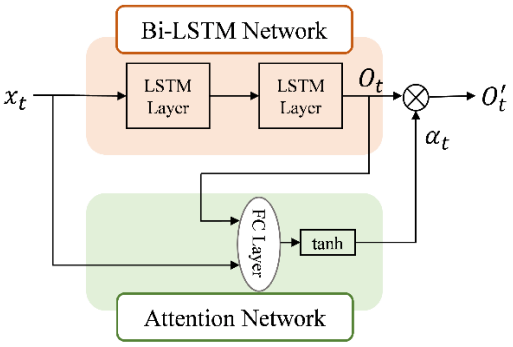


**FIGURE 5.** The graphical illustration of the attention model.

on discriminative frames with a soft attention mechanism that can automatically measure the importance of different frames.

As mentioned previously, for consecutive $T$ frames, the multi-layer bidirectional LSTM learns spatiotemporal information and outputs fire smoke related representation $O = \{O_1, O_2, \ldots, O_T\}$. The illustration of the spatial attention network is shown in Figure 5. At each time step $t$, the scores $s_t$ for indicating the importance of the $T$ frames are jointly obtained as

$$s_t = U_s tanh(W_{xs}X_t + W_{os}O_t + b_s) + b_{us} \tag{11}$$

where $U_s$, $W_{xs}$, $W_{os}$ are the weight matrices learned from the network and $b_s$, $b_{us}$ are bias vectors. $X_t$ is the spatial features extracted by CNN. $O_t$ is the spatiotemporal information extracted by Bi-LSTM. For the $k$th frame, the importance value is computed as

$$\alpha_t = \frac{exp\,(s_t)}{\sum_{i=1}^{T} exp\,(s_i)} \tag{12}$$

which is a normalization of the scores. Among the sequences, the larger the score, the more important this frame is for determining the type of classes. We regard importance values as attention weights. Instead of assigning equal degrees of importance to all the spatiotemporal information $O_t$, the final output of the attention network is modulated to $O_t^{'} = \alpha_t \odot O_t$. Finally, we concatenate all the time step output of attention network and add a softmax layer on top of the model for classification.

## IV. EXPERIMENTS

In this section, we will introduce the experimental setting in detail. Then we design several groups of experiments to measure the performance of proposed ABi-LSTM. Finally, we test the computational efficiency of the proposed framework.

### A. DATASET

There is currently no large scale forest fire smoke dataset for algorithmic train and test. We build a large-scale forest fire smoke video dataset with Nanjing Enbo Technology Company Ltd. We collect a large number of real early forest fire video to create our dataset, all videos were captured from forest fire monitoring system with an image size of $1920 \times 1080$.

Considering that dynamic feature is one of the essential features of smoke. In this paper, the foreground detection algorithm is used for the candidate patch proposal. After comparing the performance and stability of some foreground detection algorithms, the ViBe [32] background subtraction method is selected to detect the candidate patch. When the number of individual foreground target pixels exceeds a threshold (50 in this paper), the area in which the foreground target is located is considered to be a suspected target. The $299 \times 299$ image sequence centered on the moving target is fed to ABi-LSTM. The top half of the Figure 6 is the raw video sequence, and the bottom half is the foreground map obtained by VIBE.

The sequence sample is 5 frames per second, with a total length of 20 frames. The total number of sequences is 2000, including 1000 smoke containing sequences and 1000 non-smoke sequences. For purpose of training and testing, the

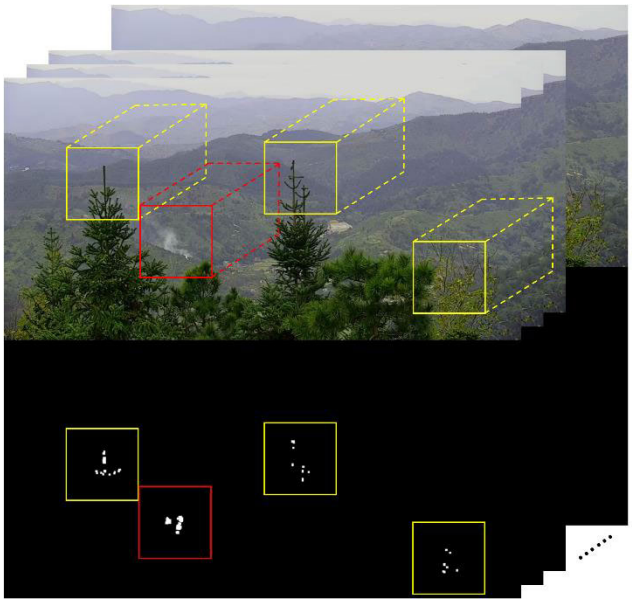

**FIGURE 6.** Using background subtraction technique to get the moving targeted region. Positive sequences are highlighted in red boxes, negative sequences are highlighted in yellow boxes.

**TABLE 1.** Dataset for experiment.

| Type | # of sets | # of frames | Purpose |
|---|---|---|---|
| Positive | 800 | 16000 | Training |
| Negative | 800 | 16000 | Training |
| Positive | 200 | 4000 | Testing |
| Negative | 200 | 4000 | Testing |

dataset is split into training and test sets, with an 80-20 split. The details of dataset are described in Table 1.

### B. IMPLEMENTATION DETAILS

Experiments were conducted on a personal computer with CPU of Intel Core i5-6500 and GPU of NVIDIA GTX1080. The proposed ABi-LSTM architecture is implemented on the TensorFlow framework.

In most of the literature, researchers normally computed accuracy based at patch-level because there is little test data. However, we evaluate the accuracy based on suquence-level evaluation that is smoke and non-smoke sequence classification accuracy in our work. The proposed ABi-LSTM framework is trained stage by stage.

In the first stage, we use Adam optimizer for Inception V3 network training. Instead of randomly initializing the weights, we use the pre-trained Inception V3 model on ImageNet to finetune, with learning rate of 0.00001, batch size of 32, input size of $3 \times 299 \times 299$, and train epoch of 30. Since our forest fire smoke recognition task is different from the ImageNet, we define a new top-level classifier on the basis of Inception V3 neural network by adding a fully connected layer. The newly stacked fully connected layer uses relu as the activation function and uses softmax for classification. In training phase, we chose to train only the top 2 inception blocks and newly stacked layer, and freeze the other 172 layers.

In the second stage, the output of the "avg_pool" layer in Inception V3 are extracted as spatial feature for each frame. The learned spatial feature and sequence label are fed to train the subsequent model. We use RMSprop optimizer for ABi-LSTM network training, with learning rate of 0.00001, batch size of 32, input size of $2048 \times 20$, and train epoch of 50.

### C. RESULTS AND COMPARISONS

In this section, we first introduce the evaluation protocol including statistical measures. Then we evaluate the performance our ABi-LSTM method with other methods. Thirdly, we do the ablation experiments of each sub-model of the proposed ABi-LSTM.

#### 1) EVALUATION PROTOCOL

For binary classification of image patch sequence, the sequence can be divided into true positive (TP), false positive (FP), true negative (TN), false negative (FN) four groups based on its combination of true class and predicted class. The predicted class is the output of the ABi-SLTM. The specific classification is as follows:

- True Positive (TP): Correctly classified as the smoke sequence
- True Negative (TN): Correctly classified as the non-smoke sequence
- False Positive (FP): Incorrectly classified as the smoke sequence
- False Negative (FN): Incorrectly classified as the non-smoke sequence

Performance of binary classifier are usually evaluated by the following widely used statistical measures: true positive rate (TPR), true negative rate (TNR) and Accuracy Rate (AR). The relative number of TP with respect to the overall number of positives is called the true positive rate (TPR), which is also known as sensitivity. The true negative rate (TNR) measures the proportion of actual negatives that are correctly identified as such. Another, Accuracy Rate (AR) is an overall measure for the relative number of correct classifications of both positives and negatives, which can be used to compare the overall performance of the different algorithms. Mathematically, these statistical measures can be expressed as:

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

$$TNR = \frac{TN}{TN + FP} \tag{14}$$

$$AR = \frac{TP + TN}{TP + FN + TN + FP} \tag{15}$$

In our ABi-LSTM model, the cross entropy loss function layer is the end with two parts: the predicted probability value $q_i$ and the true label $p_i$. For each sequence $x$, the probability of the output $y = 1$ is given by $q_{y=1} = \hat{y}$, Similarly, the probability of the output $y = 0$ is simply given by

**FIGURE 7.** Smoke sequences used in our method.

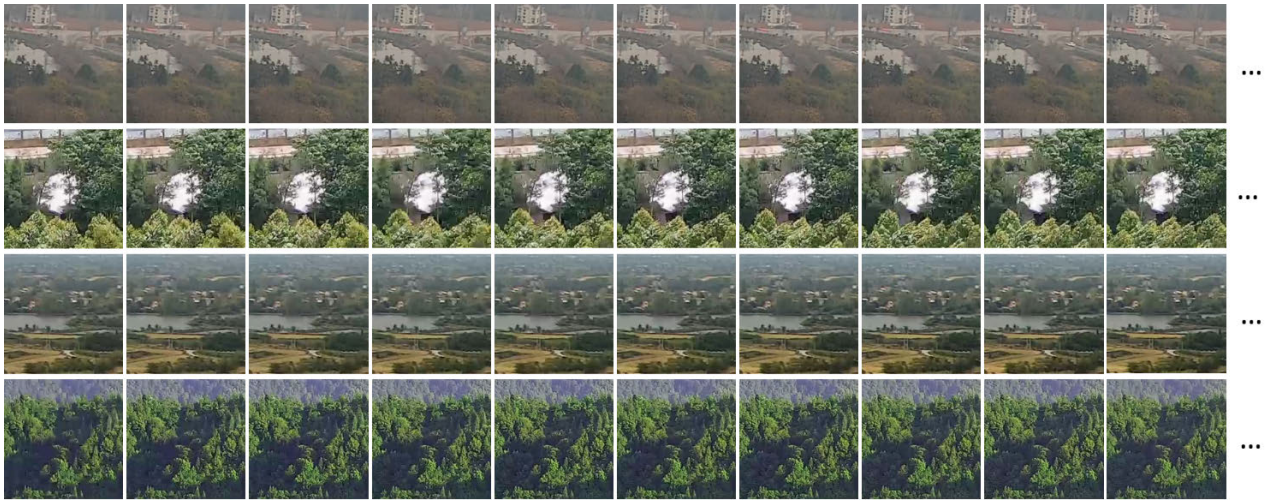

**FIGURE 8.** Non-smoke sequences used in our method.

$q_{y=0} = 1 - \hat{y}$. The true probabilities can be expressed similarly as $p_{y=1} = y$ and $p_{y=0} = 1 - y$. The loss function for the example is formulated as:

$$L(p, q) = -\sum_i p_i log q_i = -y log\hat{y} - (1-y)\, log\left(1-\hat{y}\right) \quad (16)$$

### 2) EXPERIMENTS RESULTS

We first used the test sets to find the optimal setting of our approach: learning rate, number of LSTM layers, number of LSTM hidden units, and so on.

To show the superiority of the proposed ABi-LSTM, we compare our method with Inception V3 [23], CNN+MLP, 3DCNN [20], TSN [35], ECO [36]and three common smoke and fire recognition methods [29], [30], [31]. Table 2 shows the comparison results of different methods and parameters on our dataset. In order to analyze the influence of parameters on the accuracy and complexity of the model, we compared the experimental results of CNN+MLP and ABi-LSTM under different parameters. The x in CNN-MLP-x indicates the number of hidden cells in the MLP. Similarly, ABi-LSTM-x indicates the number of bidirectional LSTM cells. The input to the other models is

**TABLE 2.** Comparison with other method on our dataset.

| Methods | TPR | TNR | AR | Runtime for sequence |
|---|---|---|---|---|
| Method in [29] | 88.3% | 89.1% | 88.7% | - |
| Method in [30] | 88.9% | 89.7% | 89.3% | - |
| Method in [31] | 91.1% | 91.9% | 91.5% | - |
| 3DCNN [20] | 96.3% | 96.1% | 96.2 % | 0.5465s |
| ECO [36] | 93.8% | 94.4% | 94.1% | 0.0861s |
| TSN-Inception V3 | 97.1% | 96.7% | 96.9% | 0.0741s |
| CNN-MLP-128 | 95.1% | 95.3% | 95.2% | 0.0748s |
| CNN-MLP-256 | 94.8% | 95.4% | 95.1% | 0.0749s |
| CNN-MLP-512 | 95.0% | 95.6% | 95.3% | 0.0752s |
| Inception V3 [23] | 93.2% | 93.6% | 93.4% | 0.0738s |
| ABi-LSTM-64 | 97.5% | 98.0% | 97.8% | 0.0964s |

**FIGURE 9.** Results of the forest fire smoke monitoring system. The detected smoke was marked in red boxes.

**TABLE 3.** Confusion matrix for the classification of smoke and non-smoke based on the ABi-LSTM-64.

| | | Output | | |
|---|---|---|---|---|
| | | Smoke sequence | Non-smoke sequence | |
| Truth | Smoke sequence | 195<br>48.8% | 5<br>1.3% | 97.5%<br>2.5% |
| | Non-smoke sequence | 4<br>1.0% | 196<br>49.0% | 98.0%<br>2.0% |
| | | 98.0%<br>2.0% | 97.5%<br>2.5% | 97.8%<br>2.2% |

chronological patch sequences, except that the input to the Inception V3 is single patches. And we report the average runtime required to process 20 frame sequences in Table 2.

As shown in Table 2, the ABi-LSTM framework achieves the total accuracy of 97.8% with true positive rate 97.5% and true negative rate 98.0%. From Table 2, we can see that the results of our proposed ABi-LSTM outperform 4.4% than image-based Inception V3 model. The comparison results prove that the ABi-LSTM is optimal for sequence-based forest fire smoke recognition.

For clarity, the confusion matrix of ABi-LSTM is shown in Table 3. Furthermore, we conduct an ablation study to evaluate the performance of each sub-model of the proposed ABi-LSTM. In this research, we conduct three models for comparison:

- Inception V3 is a single frame image model, and its experimental results are mentioned in Table 2, which is considered as baseline in ablation experiments.
- Uni-directional LSTM-x is a single-direction LSTM, in which x represents the number of hidden units. Uni-directional LSTM consists of two sub-model: the spatial features extraction network and the uni-directional LSTM network.
- Bi-LSTM consists of two sub-model: the spatial features extraction network and the Bidirectional LSTM network. The input patches are fed into the Bi-LSTM network one by one.
- ABi-LSTM consists of all the three sub-model: the spatial features extraction network, the Bidirectional LSTM network, and the temporal attention subnetwork. The input patches are fed into the ABi-LSTM network one by one.

As shown in the Table 4, the Bi-LSTM network improves the accuracy of the Inception V3 imaged-based model by 2.1%, and the temporal attention subnetwork improves the accuracy of the Bi-LSTM model by 2.3%. The ablation experiments justify our initial design idea.

**TABLE 4. The ablation analysis of the ABi-LSTM.**

| Methods | TPR | TNR | AR |
|---|---|---|---|
| Inception V3[23] | 93.2% | 93.6% | 93.4% |
| Uni-directional LSTM-32 | 94.5% | 95.1% | 94.8% |
| Uni-directional LSTM-64 | 94.8% | 95.4% | 95.1% |
| Uni-directional LSTM-128 | 94.7% | 95.3% | 95.0% |
| Bi-directional LSTM-32 | 94.8% | 95.6% | 95.2% |
| Bi-directional LSTM-64 | 95.2% | 95.8% | 95.5% |
| Bi-directional LSTM-128 | 94.9% | 95.7% | 95.3% |
| ABi-LSTM-64 | 97.5% | 98.0% | 97.8% |

**TABLE 5. Average complexity comparisons of an image.**

| Methods | GFLOPs | Million Parameters | AR |
|---|---|---|---|
| 3DCNN [20] | - | 178.7 | 96.2% |
| TSN-Inception V3 | 5.746 | 23.8 | 96.9% |
| CNN-MLP-128 | 5.748 | 26.4 | 95.2% |
| CNN-MLP-256 | 5.749 | 34.2 | 95.1% |
| CNN-MLP-512 | 5.752 | 44.1 | 95.3% |
| Inception V3 [23] | 5.746 | 23.8 | 93.4% |
| ABi-LSTM-64 | 5.761 | 26.1 | 97.8% |

The proposed model can be deployed easily, which can be used to recognize the suspected smoke patches in practical application. Figure 9 shows the recognition results of proposed model.

## V. CONCLUSION

In this paper, we propose an attention enhanced bidirectional LSTM network (ABi-LSTM) for early forest smoke recognition. Specifically, the proposed approach can be summarized as three parts:

a) an Inception V3 network which is used to extract spatial features from smoke candidate patch step by step; b) Bi-LSTM model which is designed to extract temporal features from forward and backward order by feed spatial feature of single patch; c) attention network is employed to optimize classification process with a soft attention mechanism that can automatically measure the importance of different frames. Extensive experiment results show that the proposed ABi-LSTM framework obtains higher accuracy in early forest fire smoke recognition compared with other methods. Moreover, ablation study is conducted to evaluate the performance of each sub-model in ABi-LSTM.

The proposed ABi-LSTM has been inspired by the attention mechanism in neural machine translation, which can adaptively focus on discriminative frames. As a result, this framework may be suitable for early forest fire smoke detection. An interesting question is whether attention mechanism can be used in a single frame image to enable the model to learn more discriminatory spatial information. This will be investigated in the future.

## REFERENCES


[1] H. Tian, W. Li, P. O. Ogunbona, and L. Wang, "Detection and separation of smoke from single image frames," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1164–1177, Mar. 2018.

[2] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A deep normalization and convolutional neural network for image smoke detection," *IEEE Access*, vol. 5, pp. 18429–18438, 2017.

[3] F. Yuan, X. Xia, J. Shi, H. Li, and G. Li, "Non-linear dimensionality reduction and Gaussian process based classification method for smoke detection," *IEEE Access*, vol. 5, pp. 6833–6841, 2017.

[4] K. Dimitropoulos, P. Barmpoutis, and N. Grammalidis, "Higher order linear dynamical systems for smoke detection in video surveillance applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 1143–1154, May 2017.

[5] G. Lin, Y. Zhang, G. Xu, and Q. Zhang, "Smoke detection on video sequences using 3D convolutional neural networks," *Fire Technol.*, vol. 55, no. 5, pp. 1827–1847, Sep. 2019.

[6] A. E. Çetin, K. Dimitropoulos, B. Gouverneur, N. Grammalidis, O. Günay, Y. H. Habiboğlu, B. U. Töreyin, and S. Verstockt, "Video fire detection—Review," *Digit. Signal Process.*, vol. 23, no. 6, pp. 1827–1843, 2013.

[7] J. A. Ojo and J. A. Oladosu, "Video-based smoke detection algorithms: A chronological survey," *Comput. Eng. Intell. Syst.*, vol. 5, no. 7, pp. 38–50, 2014.

[8] C. Long, J. Zhao, S. Han, L. Xiong, Z. Yuan, J. Huang, and W. Gao, "Transmission: A new feature for computer vision based smoke detection," in *Artificial Intelligence and Computational Intelligence*, vol. 6319, F. L. Wang, H. Deng, Y. Gao, and J. Lei, Eds. Berlin, Germany: Springer, 2010, pp. 389–396.

[9] F. Yuan, "A double mapping framework for extraction of shape-invariant features based on multi-scale partitions with AdaBoost for video smoke detection," *Pattern Recognit.*, vol. 45, no. 12, pp. 4326–4336, 2012.

[10] H. Tian, W. Li, L. Wang, and P. Ogunbona, "Smoke detection in video: An image separation approach," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 192–209, Jan. 2014.

[11] H. Tian, W. Li, P. Ogunbona, and L. Wang, "Single image smoke detection," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 87–101.

[12] L. Millan-Garcia, G. Sanchez-Perez, M. Nakano, K. Toscano-Medina, H. Perez-Meana, and L. Rojas-Cardenas, "An early fire detection algorithm using IP cameras," *Sensors*, vol. 12, no. 5, pp. 5670–5686, May 2012.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, 2012, pp. 1097–1105.

[14] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, 2015, pp. 577–585.

[15] Y. Goldberg, "Neural network methods for natural language processing," *Synth. Lectures Hum. Lang. Technol.*, vol. 10, no. 1, pp. 1–309, 2017.

[16] Y. Zhao, J. Ma, X. Li, and J. Zhang, "Saliency detection and deep learning-based wildfire identification in UAV imagery," *Sensors*, vol. 18, no. 3, p. 712, Feb. 2018.

[17] Y. Liu, W. Qin, K. Liu, F. Zhang, and Z. Xiao, "A dual convolution network using dark channel prior for image smoke classification," *IEEE Access*, vol. 7, pp. 60697–60706, 2019.

[18] G. Xu, Q. Zhang, D. Liu, G. Lin, J. Wang, and Y. Zhang, "Adversarial adaptation from synthesis to reality in fast detector for smoke detection," *IEEE Access*, vol. 7, pp. 29471–29483, 2019.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, 2015, pp. 91–99.

[20] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.

[21] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," Dec. 2017, *arXiv:1801.01078*. [Online]. Available: https://arxiv.org/abs/1801.01078

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[24] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.
[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[26] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 4263–4270.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2015, pp. 1–9.
[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Feb. 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167
[29] F. Yuan, "Video-based smoke detection with histogram sequence of LBP and LBPV pyramids," *Fire Saf. J.*, vol. 46, no. 3, pp. 132–139, Apr. 2011.
[30] H. Tian, W. Li, P. Ogunbona, D. T. Nguyen, and C. Zhan, "Smoke detection in videos using non-redundant local binary pattern-based features," in *Proc. IEEE 13th Int. Workshop Multimedia Signal Process.*, Oct. 2011, pp. 1–4.
[31] M. Favorskaya, A. Pyataeva, and A. Popov, "Verification of smoke detection in video sequences based on spatio-temporal local binary patterns," *Procedia Comput. Sci.*, vol. 60, pp. 671–680, Jan. 2015.
[32] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, p. 1330, 2017.
[33] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.
[34] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.
[35] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 20–36.
[36] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis.*, vol. 11206, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 713–730.

**YICHAO CAO** received the B.S. degree from the School of Electrical and Information Engineering, Jiangsu University, China, in 2015, and the M.S. degree in automation from Southeast University, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Automation. His research interests include image processing, deep learning, and computer vision.

**FENG YANG** graduated from the School of Electrical and Information Engineering, Jiangsu University of Technology, China, in 2017. She is currently pursuing the master's degree with the School of Instrument Science Engineering, Southeast University, China. Her research interests include image processing, computer vision, and simultaneous localization and mapping.

**QINGFEI TANG** received the B.S. degree from the Chongqing University of Science and Technology, Chongqing, China, in 2015, and the M.S. degree from Northeastern University, Shenyang, China, in 2018. He is currently an Algorithm Researcher with Nanjing Enbo Technology Company Ltd. His research interests include image retrieval, image detection, and video classification.

**XIAOBO LU** received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, the M.S. degree from Southeast University, Nanjing, China, and the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics. He did a Postdoctoral research with the Chien-Shiung Wu Laboratory, Southeast University, from 1998 to 2000, where he is currently a Professor with the School of Automation and the Deputy Director of the Detection Technology and Automation Research Institute. He is the coauthor of the book *An Introduction to the Intelligent Transportation Systems* (Beijing, China Communications: 2008). His research interests include image processing, signal processing, pattern recognition, and computer vision. He has received many research awards such as the First Prize of the Natural Science Award from the Ministry of Education of China and a prize of the Science and Technology Award of Jiangsu Province.