Computer Science and Information Systems 20(4):1707–1728 https://doi.org/10.2298/CSIS221125027L

Generative Adversarial Network Based on LSTM and Convolutional Block Attention Module for Industrial Smoke Image Recognition

Dahai Li[1,⋆], Rui Yang[1], and Su Chen[2]

[1] School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology
Zhengzhou 450064, China
ldadahai@163.com
yangrui2233@163.com

[2] Department of Mechanical and Electrical Engineering, Henan Vocational College of Water Conservancy and Environment
Zhengzhou 450002, China
chensssu@163.com

Abstract. The industrial smoke scene is complex and diverse, and the cost of labeling a large number of smoke data is too high. Under the existing conditions, it is very challenging to efficiently use a large number of existing scene annotation data and network models to complete the image classification and recognition task in the industrial smoke scene. Traditional deep learn-based networks can be directly and efficiently applied to normal scene classification, but there will be a large loss of accuracy in industrial smoke scene. Therefore, we propose a novel generative adversarial network based on LSTM and convolutional block attention module for industrial smoke image recognition. In this paper, a low-cost data enhancement method is used to effectively reduce the difference in the pixel field of the image. The smoke image is input into the LSTM in generator and encoded as a hidden layer vector. This hidden layer vector is then entered into the discriminator. Meanwhile, a convolutional block attention module is integrated into the discriminator to improve the feature self-extraction ability of the discriminator model, so as to improve the performance of the whole smoke image recognition network. Experiments are carried out on real diversified industrial smoke scene data, and the results show that the proposed method achieves better image classification and recognition effect. In particular, the F scores are all above 89%, which is the best among all the results.

Keywords: industrial smoke image recognition, generative adversarial network, LSTM, convolutional block attention module, data enhancement.

1. Introduction

Bad weather conditions will have an impact on our production and life. If the weather can not be accurately identified, timely judged, it will lead to the threaten to human life and property. The weather can change very quickly sometimes. In particular, fog, rain, snow, smoke, sand and other extreme weather will affect all walks of life [1,2]. Industrial smoke

⋆ Corresponding author

weather, as an important scene, widely exists in the daily life of big cities, and various tasks for the smoke scene have become very urgent, and gradually become one of the hot spots in computer vision area. In order to successfully complete the classification and recognition task of smoke and dust image, it is necessary to provide a sufficient number of manually labeled images for each specific smoke and dust scene in advance [3]. However, in practice, it costs a lot of cost to acquire enough diverse smoke labeled data samples. With the emergence of large amount of labeled training data and large amount of computing resources, deep neural network can achieve satisfactory performance. These data and deep neural network models can be directly and widely applied to image classification in most scenarios. However, when dealing with the classification of smoke scene, there will be a huge classification error. In order to make use of the existing conditions for image recognition of smoke scene [4], it is necessary to solve the problem of data distribution difference. However, the data distribution of smoke scene is more different from that of other conventional situations. At the same time, data with such a difference in distribution also have a larger domain shift. Convolutional neural networks are difficult to learn invariance features of data.

In view of the above problems, existing research methods provide some solutions, which are mainly divided into two categories. On the one hand, based on the existing convolutional neural network, combined with data enhancement methods, for example, PM-SNet (Robust haze removal based on patch map for single images) [5], DCPDN (densely connected pyramid dehazing network) [6] remove haze to reduce the distribution difference between haze and normal scenario data. However, most fog removal algorithms require a large number of paired images for training. Due to the particularity of fog environment, fog scene images in the training data are mostly synthetic, so it is difficult to reduce the distribution difference in the pixel field. According to Huang et al, [7] data enhancement using fog removal algorithm has no significant gain in fog classification and recognition task in real scene. And collecting and generating more data will increase training time and cost. In addition, atmospheric degradation model is used to train the data. However, this method is difficult to apply to the true diversity of fog scenes.

On the other hand, it is difficult to extract invariable features in feature domain only by using data enhancement-based deep neural network due to different distribution between different data sets. The domain adaptive algorithm can reduce the domain offset and learn invariance features of different data. Existing domain adaptive algorithms use different feature alignment methods to align features extracted from convolutional neural networks, and some methods reduce domain offset based on maximum mean difference. For example, DANN (Domain Adaptive Neural Network) [8] used maximum mean difference measurement to reduce feature distribution between two hidden layers by samples extracted from different domains. Deep domain confusion (DDC) [9] used an adaptation layer to connect two parallel CNN (convolutional neural Network [10,11]) models, and utilized MMD (Maximum Mean Discrepancy) to define the loss function and narrow the domain offset of source domain and target domain. Based on CORAL method, the nonlinear transformation method was used to align the second-order statistical features of source domain and target domain distribution. Another adversarial approach, such as DANN-A (Domain-Adversarial Training of Neural Networks) [12], used adversarial networks to obtain domain-invariant features. Most of the existing domain adaptive algorithms focus on the edge distribution of aligning single feature. A single feature only represents the

characteristic information of a small part of data. Some adaptive algorithms will appear unstable training and poor robustness when there is a large field offset.

To solve the above problems, Zhao et al. [13] used CycleGAN (Unpaired image-to-image translation using Cycle-consistent Adversarial Networks) for data enhancement. On the one hand, data enhancement can be carried out in the normal scenario (source domain) to generate labeled data with a style similar to that in the fog scenario. Cycle-GAN used conventional data and foggy scene data (target domain) for antagonistic training, which could reduce the data distribution difference between them in pixel domain. Since Cycle-GAN did not need to use paired images and additional data, which greatly saved training time and training cost.

Chen et al. [14] proposed an effective adaptive network, namely multi-scale featuresand multi-adversarial network (MFMAN) and proposed an IAM (Inception adversarial) network to fully align data distribution in feature domain and reduce field offset (perceptual adversarial module). After feature extraction by convolutional neural network, IAM was used to extract multiple features of different scales to replace a single feature. Feature information extracted by convolutional neural network was fully utilized to form adversarial by multiple domain discriminators to reduce feature distribution differences. In order to make use of the complex multi-mode structure, Karnewar et al. [15] used multiple discriminators to learn the multi-mode structure of data distribution on the category, which increased the discriminant ability of the category, thus increasing the positive migration and reducing the negative migration. Features could reduce domain differences in each category and learn domain invariance features.

Our main contributions are as follows.

- We set up a new industrial smoke data set. a low-cost data enhancement method is used to effectively reduce the difference in the pixel field of the image.
- The smoke image is input into the LSTM in generator and encoded as a hidden layer vector. This hidden layer vector is then entered into the discriminator.
- Meanwhile, a convolutional block attention module is integrated into the discriminator to improve the feature self-extraction ability of the discriminator model, so as to improve the performance of the whole smoke image recognition network.

The remaining paper is shown as follows: Section 2 presents the related works in detail. Section 3 discusses the data sets used in this paper. Generator model based on encoder-decoder framework and Discriminator are stated for the proposed method in section 4 and section 5 respectively. In Section 6, experiments are conducted and analyzed. The paper is concluded in Section 7.

2. Related Works

Image segmentation is a basic computer vision technology. Its purpose is to divide an image into visually meaningful areas with similar properties, such as mountains and people, according to the features of color, texture, gray level and motion. Because of its fundamental role in the field of computer vision, image segmentation plays an important role in image understanding, content-based video compression and coding, content-based image and video retrieval, etc [16].

In recent years, scholars have carried out a large number of relevant researches on this task, which can be divided into traditional digital image processing method and deep learning method according to different research methods.

In image segmentation method, threshold technique [17] is a simple and effective method. The basic idea of threshold segmentation is as follows. First, it obtains the gray histogram corresponding to the image, and then determines a gray threshold value according to the histogram information as the basis of segmentation. Finally, pixel points in the image whose gray value is greater than the threshold value are classified into one class (such as the target region), while pixel points in the image whose gray value is less than or equal to the threshold value are classified into another class (such as the background region).

At present, there are two common threshold segmentation methods: bimodal method and maximum variance method [18]. Bimodal method means that the background and the object of concern in the image form a crest respectively on the gray histogram of the image, that is, the region corresponds to the crest one by one. In this way, as long as the gray value corresponding to the trough between the two wave peaks is selected as the threshold, the object and the background can be separated well. Bimodal method is applicable to the grayscale distribution of each object in the image is relatively regular, and many soot images do not conform to this situation. The maximum variance method can automatically select the threshold value without manual parameter setting, and the selected threshold value can obtain better segmentation effect in most cases.

The threshold segmentation algorithm of the maximum variance method is described as follows. It is assumed that the image to be segmented contains two types of regions: Region 1 and Region 2. t is set as the threshold for separating two regions, and according to histogram statistics, the area ratio of region 1 and region 2 separated by t to the whole image is θ_1 and θ_2:

$$\theta_1 = \sum_{j=0}^{t} \frac{n_j}{n} \tag{1}$$

$$\theta_2 = \sum_{j=t+1}^{G-1} \frac{n_j}{n} \tag{2}$$

Where n is the area of the whole image. n_j is the corresponding area when the gray level is j. The average gray level of Region 1, Region 2 and the whole image is μ_1, μ_2 and μ:

$$\mu_1 = \frac{1}{\theta_1} \sum_{j=0}^{t} (f_i \times \frac{n_j}{n}) \tag{3}$$

$$\mu_2 = \frac{1}{\theta_1} \sum_{j=t+1}^{G-1} (f_i \times \frac{n_j}{n}) \tag{4}$$

$$\mu = \sum_{j=0}^{G-1} (f_i \times \frac{n_j}{n}) \tag{5}$$

Where f is the gray value corresponding to gray j.

The variance between the two regions after the image is segmented by the threshold value is shown as follows:

$$\sigma_B^2(t) = \theta_1(\mu_1 - \mu) + \theta_2(\mu_2 - \mu) \tag{6}$$

When the variance between the two divided regions reaches the maximum, it is considered that the two regions have reached the best separation state, and the corresponding threshold t is the best threshold T to be found.

$$T = max\lfloor\sigma_B^2(t)\rfloor \tag{7}$$

At present, deep networks such as convolutional neural networks (CNN) are widely used in image classification, object detection and other tasks, and have achieved higher accuracy than traditional digital image processing methods. Although it requires a large amount of data to train the deep network, it has the advantage of better adaptability and provides an end-to-end solution, that is, the data from the input end can be directly obtained from the output end, avoiding multiple steps such as image preprocessing and result repair, which are often used in the process of digital image processing methods to divide smoke and dust. The method of using deep network model to detect soot in image includes target detection and semantic segmentation.

Object detection is a detection method based on candidate region. For example, Teng et al. [19] applied Faster R-CNN to video smoke detection, avoiding the complicated artificial feature extraction processing in traditional smoke detection methods. Firstly, the training data set was expanded by stitching smoke for forest and other scene images, so as to train the classifier to carry out target detection of smoke. Wu et al. [20] proposed a smoke detection method combining mixed Gaussian model and YOLOV2 network. Firstly, the mixed Gaussian algorithm was used to determine the approximate soot region, and on this basis, the YOLO network was trained to predict the final soot region. The result of target detection was to determine the soot area in the image with a rectangular frame, which leaded to the inevitable inclusion of a large number of background areas in the result, which would affect the subsequent calculation result of pollution level based on the Ringelmann smoke blackness level.

The semantic segmentation method classifies each pixel so as to achieve pixel level segmentation of image target area. Li et al. [21] used the fully convolutional network model for target segmentation of soot image. By manually marking the smoke region in the smoke image to make a data set, and then training the fully convolutional network, the trained model could predict the smoke image and give pixel-level smoke segmentation results. Wang et al. [22] designed a network model containing two branches of rough segmentation and fine segmentation, both adopted a fully convolutional network with encoding and decoding structure. Li et al. [23] adopted a conditional generation adversarial network model to segment smoke regions in continuous video frames. Although the method based on the fully convolutional network has achieved a high accuracy, the segmentation accuracy in complex scenes needs to be improved. For example, the interference of clouds with high similarity to soot in the background will affect the segmentation, and it is often a problem to judge a part of the area belonging to clouds as soot. Therefore, we propose a novel generative adversarial network based on LSTM and convolutional block attention module for industrial smoke image recognition.

3. Data Sets

According to experimental requirements, real industrial smoke scene classification data set and normal scene classification data set are needed for training and testing. Currently, some larger public datasets, such as Caltech-256, PASCALVOC (Pattern Analysis, Statistical Modeling and Computational Learning Visual Object Classes) and ImageNet are mainly normal image scenarios. Image data sets in normal scenarios used in this paper are mainly from these large open source datasets, and 12 categories are extracted to form a Normal-12 set. This data is usually easy to obtain and does not require annotation. Smoke images are mainly obtained from images and videos on the Internet, they are manually classified and labeled, and named as Smoke-12 set. Smoke image data sets cover most industrial scenes. Normal-12 has 4800 images including 3600 images for training set and 1200 images for test set. Smoke-12 has 2400 images as target domain data. Detailed information of images is shown in table 1, and Figure 1 shows partial images in the dataset.

Table 1. Samples in the data sets

Name	Normal-12 train	Normal-12 test	Smoke-12
truck	270	95	170
tree	300	103	200
train	307	100	220
streetlamp	290	95	180
people	300	100	195
car	301	100	210
bus	282	92	190
building	306	102	200
bridge	315	104	215
boat	300	100	200
bird	301	100	182
plane	328	111	238
total	3600	1200	2400

Cycle-GAN algorithm [24] is used for data enhancement, as shown in Figure 2. The structure consists of a pair of generator networks and a pair of discriminator networks. A group of unaligned images are input for training to learn the pixel-level mapping between input images and output images, and the images are converted from source domain A to target domain B. The goal is to learn a map G_AB: $A \rightarrow B$ and use the discriminator D_B to narrow the distribution between Fake_B and B. Similarly, in G_BA: $B \rightarrow A$, the discriminator D_A is used to narrow the distribution of Fake_A and A, and a cyclic consistency loss is introduced to enforce $F(G(A)) \approx B$. When cycle-GAN is used for processing data enhancement, Normal-12 is used as source domain A, Smoke-12 serves as the target domain B. Through such a cyclic adversarial network structure, labeled data similar to the smoke scene style, namely Fake_B, can be obtained, and at the same time, the distribution differences of data in the two scenes can be initially reduced in the pixel domain.

Fig. 1. Some samples in data sets

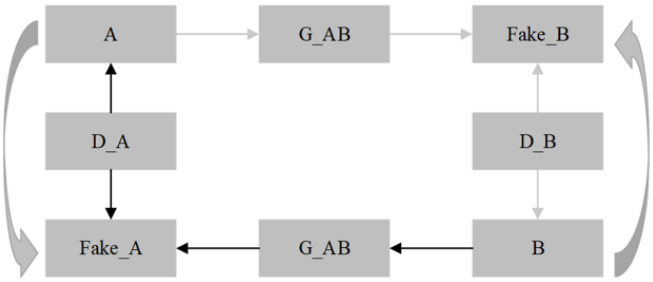

Fig. 2. Cycle-consistent adversarial network structure

4. Generator Model Based on Encoder-Decoder Framework

This section introduces the generator model G of Encoder-Decoder framework based on LSTM in detail. The purpose of the model is to learn the distribution function P_g of the generator on sample data. The distribution function of real samples is P_{data}. First, it inputs image S into LSTM (named EnLSTM) in Encoder stage, and obtains hidden layer vector with fixed dimension after encoding. Then, the hidden layer vector is input into LSTM (named DeLSTM) in Decoder stage [25]. The hidden layer vector at t moment is generated by combining the output at $t-1$ moment, and the generated sequence y is finally decoded. The mapping process is expressed as $G(S,\theta)$. θ is the parameter of LSTM. The generator model structure is shown in Figure 3. In the following, we give the detailed Encoder process and Decoder process in Generator.

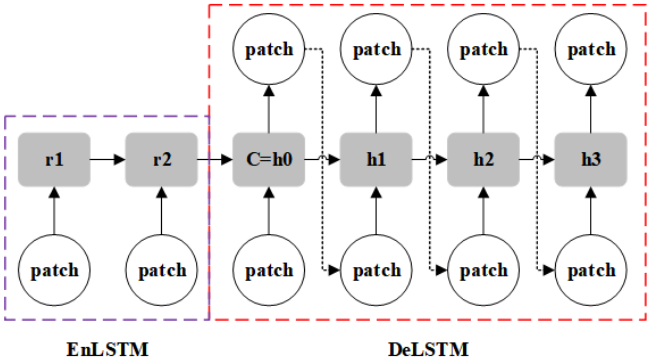

Fig. 3. Generator model

4.1. Encoder

As shown in Figure 3, the left purple box is the Encoder process. The $i-th$ image slice in S is denoted as s_i, slice sequence $S = s_1, s_2, \cdots, s_t$ is defined as EnSen. Then, the input is sequentially into EnLSTM, and a slice vector is input for each time step t. The hidden layer vector r_t at the current moment t is determined by the hidden layer vector r_{t-1} at the $t-1$ moment and the input s_t at the current moment. LSTM has the ability to remove or add information to a hidden state. LSTM includes three gate structures, namely, forgetting gate, input gate and output gate. First, the forgetting gate layer decides to discard some information:

$$f_t = \sigma(W_f \cdot [r_{t-1}, s_t] + b_f) \tag{8}$$

Where W_f is the weight that decides to forget the information and b_f is the bias value.

The input gate determines what information is to be updated. This step has two parts: First, the Sigmoid layer, called the input gate layer, determines what values are to be

updated. Then, it creates a new candidate value vector $\tilde{C}_t$ through tanh layer and adds it to the state, namely:

$$i_t = \sigma(W_i \cdot [r_{t-1}, s_t] + b_i) \tag{9}$$

$$\tilde{C}_t = tanh(W_c \cdot [r_{t-1}, s_t] + b_c) \tag{10}$$

W_i and W_c are the weights of Sigmoid layer and tanh layer respectively. b_i and b_c are the offset values. Based on the above calculation, the updated status information is:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{11}$$

Finally, the output information is determined through the output gate, and then the vector representation of the output information is obtained as shown in formula (12) and formula (13):

$$o_t = \sigma(W_o \cdot [r_{t-1}, S_t] + b_o) \tag{12}$$

$$r_t = o_t \cdot tanh(C_t) \tag{13}$$

W_o is the weight that determines the output information. b_o is the offset value. After the above coding, a hidden layer vector sequence $r = r_1 : r_t$ is obtained. When it is input to end flag (EOS), the final semantic vector r_c is generated. r_c is input to the decoder stage as the initial vector.

4.2. Decoder

The red box on the right in Figure 3 is the Decoder network structure diagram of the generator model. In this stage, the semantic vector r_c obtained in the Encoder stage is first input into DeLSTM, which is denoted as h_0 as the hidden layer vector at the initial t_1 moment. Then it is decoded to obtain the output y_0 at time t_1. At time t_1, the input of DeLSTM is the output y_1. Then the hidden layer vector h_2 at time t_2. We use the same method to obtain the each hidden layer vector at t moment. Then through the three gate operations of DeLSTM such as formulas (8)-(13), the implicit vector sequence $h = [h_0 : h_1]$ of DeLSTM is obtained. Then, the hidden layer vector at time t is input into a Softmax function to generate a probability distribution of N words in the image description dictionary C, so that each word in the dictionary at time t belongs to $n(1, \cdots, N)$ the probability of class words. For example, the dictionary size is $V :< W_{o1}^t, W_{o2}^t, \cdots, W_{oV}^t >$, so:

$$P(W_{oV}^t = n|EnSen, \theta) = \frac{exp(\omega_n^T h_t)}{\sum_{n=1}^{N} exp(\omega_n^T h_t)} \tag{14}$$

Where ω_c is the weight matrix of decoder LSTM from the hidden layer to the output sequence. These probability values can be further obtained by an Argmax function to obtain the output sequence $y_1, y_2, \cdots, y_t$, defined as DeSen. o is the parameter to generate the LSTMs in the model. This process predicts the image slice y_t of the next output according to the given semantic vector r_c and the previously generated output sequence $y_1, y_2, \cdots, y_{t-1}$, that is:

$$y_t = argmaxP(y_t|y_1, y_2, \cdots, y_{t-1}, r_c, h_t) \tag{15}$$

Equation (15) can be abbreviated as:

$$y_t = g(y_{t-1}, h_t, r_c) \tag{16}$$

Where h_t is the hidden layer vector of DeLSTM. y_{t-1} is the output of the previous moment. g is a nonlinear multilayer perceptron that generates the probability that each slice in dictionary C belongs to y_t.

5. Discriminator Based on Convolutional Block Attention Module (CBAM)

5.1. Generative Adversarial Network (GAN)

The traditional generative adversarial network (GAN) [26] consists of a generator (G) and a discriminator (D). G is to generate sample data similar to or even consistent with real samples according to the characteristics of real samples, so as to supplement the lack of sample types and achieve a balance of sample distribution. D is to supervise the quality of the generated samples and judge whether the generated samples are consistent with the characteristics of the real samples, so as to achieve the purpose of expanding the number of samples. For generator G, the data input only contains the generated sample, so the objective function of G is:

$$min_G L(G) = minE[logP(S = F_{fake}|X_{fake})] \tag{17}$$

Where S stands for data source. X_{fake} represents the sample generated by the generator. $P(S = F_{fake}|X_{fake})$ represents the probability that the generated sample belongs to the generated sample data source. E is the expectation calculation. According to Formula (10), the optimization goal of G is to make the probability of the generated sample being judged as "false sample" as small as possible. When enough samples are generated to "look like real", the optimization goal is completed. For D, data input includes both real samples and generated samples, and its objective function is shown in Equation (18).

$$min_D L(D, G) = maxE[logP(S = R_{real}|X_{real})] + E[logP(S = F_{fake}|X_{fake})] \tag{18}$$

Where X_{real} represents the real sample. $P(S = real|X_{real})$ represents the probability that the real sample belongs to the real sample data source. Its optimization objective is: whether the input is real sample or generated sample, the final result can give the correct judgment. In the process of training, GAN plays a game with each other through alternate optimization and finally achieves Nash equilibrium to complete the model training.

5.2. GAN-CBAM

The results of sample source and sample label judgment are output by D, and the performance of D determines the overall performance of GAN. Therefore, in this paper, CBAM [27] is added to the feature extraction network of D to focus on the detailed features in all kinds of transient SPM tracks, so as to improve the feature capture ability of the discriminator network and make it better calculate training errors.

CBAM can be divided into channel attention network and spatial attention network. Firstly, average pooling and maximum pooling are used to aggregate the whole channel information to generate two different spatial features F_{avg} and F_{max}. Then, the results are weighted through the multi-layer perceptron (MLP) [28] network containing a single implicit layer, and the parameter distribution weights of each channel are obtained. The calculation process is shown in Figure 4.

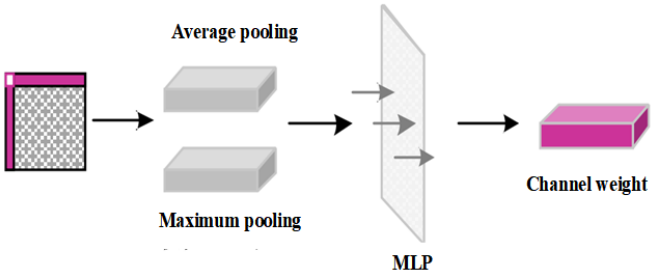

Fig. 4. Schematic diagram of channel attention

The calculation expression is shown below:

$$M_C(F) = \sigma(MLP(avgPool(F)) + MLP(maxPool(F))) \tag{19}$$

Where F represents the input convolution feature. σ stands for sigmoid function. MLP processes parameter sharing.

The output of channel attention is used as the input of spatial attention. As shown in Figure 5, maximum pooling and average pooling are performed for the data at each position in the feature matrix of each channel. And then convolution dimension reduction aggregation is performed for the two pooled results. Finally, sigmoid function is used to get spatial attention feature.

The expression of feature calculation is:

$$\begin{aligned} M_C(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^S_{avg}\cdot; F^S_{max}])) \end{aligned} \tag{20}$$

f stands for dimensionality reduction convolution. In general, 7×7 convolution kernel dimension reduction convolution calculation is adopted (the size of the eigenmatrix before and after convolution remains unchanged by adding 0).

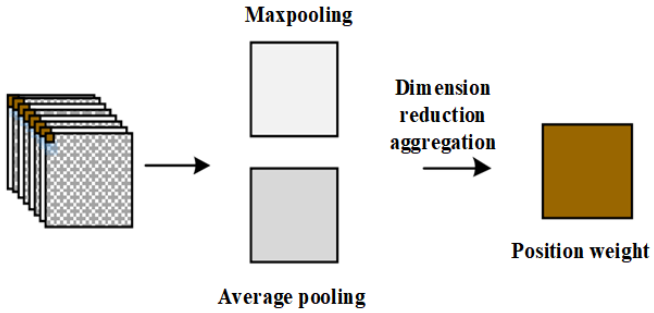

Fig. 5. Diagram of spatial attention

5.3. Loss Function and Training Process

The proposed GAN loss function in this paper consists of the global discriminator loss function, the local discriminator loss function and the tag predictor loss function, which is defined as:

$$L(\theta_f, \theta_m, \theta_y, \theta_d, \theta_d^c|_{c=1}^C) = L_y + \lambda(L_g + L_l) \tag{21}$$

Where θ_f is the backbone network parameter. θ_m is LSTM feature parameter. θ_y is tag classifier parameter. θ_d is global discriminator loss. θ_d^c is local discriminator loss. λ is the hyperparameter, which is used to determine the proportion of local loss and global loss in the total loss.

$$\theta_f \leftarrow \theta_f - \mu(\frac{\partial L_y}{\partial \theta_f} - \lambda(\frac{\partial L_g}{\partial \theta_f} - \frac{\partial L_l}{\partial \theta_f})) \tag{22}$$

$$\theta_m \leftarrow \theta_m - \mu(\frac{\partial L_y}{\partial \theta_m} - \lambda(\frac{\partial L_g}{\partial \theta_m} - \frac{\partial L_l}{\partial \theta_m})) \tag{23}$$

$$\theta_y \leftarrow \theta_y - \mu\frac{\partial L_y}{\partial \theta_y} \tag{24}$$

$$\theta_m \leftarrow \theta_m - \mu\frac{\partial L_m}{\partial \theta_m} \tag{25}$$

$$\theta_d \leftarrow \theta_d - \mu\frac{\partial L_g}{\partial \theta_d} \tag{26}$$

$$\theta_d^c \leftarrow \theta_c^c - \mu\frac{\partial L_l}{\partial \theta_d^c} \tag{27}$$

Where μ is the learning rate.

Stochastic gradient Descent (SGD) is used as the optimizer in the optimization process. However, in Formula (22), the discriminator loss is opposite to the label classifier sign loss. The discriminant loss will make the field offset of the two data bigger and bigger, the purpose of confrontation cannot be achieved. Similar some other optimizers, such

as Adagrad, Adaelta, cannot be optimized, so GRL(Gradient Reversal Layer) is added here. The network layer has no parameters related to it, so it is a normal connection layer in forward propagation and does not affect network propagation. During the back-propagation, the sign of the gradient will be changed, it is multiplied by -1, and then the back-propagation will continue. The purpose of reversing the gradient will be achieved through this network layer, so that optimization can be carried out using the optimizer.

6. Experiments and Analysis

The proposed dataset Normal-12 and Smoke-12 are used here, as shown in Figure 1 for a partial sample dataset. It contains a large number of industrial smoke scenes with different concentrations. The data set is diverse and industrial smoke images are mostly taken from common scenes in daily life. 3600 images in Normal-12 are selected as the original data. 1200 images are selected as the test set of data for training the basic classification network, and 2300 images in Smoke-12 are selected as the target data set of smoke classification and recognition.

The hardware configuration of the experimental platform is Intel(R) Xeon (R) Gold5118 CPU @ 2.30ghz, memory 256 GB, NVIDIATesla(16 GB) graphics card, software configuration is Centos7.5 using pytorch deep learning framework. Images are cropped to 256×256 pixels, and then it uses the Random Crop to randomly crop them to 224×224 pixels. Batch size is set to 128. The initial learning rate of the pre-trained backbone network is 0.001. The initial learning rate of other network layers is 0.01. The learning rate change strategy follows the existing work, the weight attenuation is 0.0005, the momentum is set as 0.9, and there are 50 training rounds in total. Based on the existing experience, grid search experiment is carried out with parameters ranging from 0 to 1. When λ is 0.3 on this data set, the convergence is faster and the accuracy is higher. The experiment mainly tests the ratio of the accurate prediction quantity of each category to the category quantity in the target domain as the category accuracy. The ratio of the number of accurate predictions of all categories to the number of smoke scene data is taken as the total accuracy.

In order to effectively evaluate the effectiveness of the proposed algorithm in smoke data set classification, the GAN is used as the basic classification network and the backbone network. In order to verify the effectiveness of adding CBAM and LSTM modules and the effectiveness of data enhancement, ablation experiments are conducted on the basic classification network for different data enhancement methods. The selected basic classification networks include GAN, MCW [29], ALSTM [30], KutralNext [31].

In order to verify that the new GAN is more effective than other image recognition algorithms in data enhancement, Normal-12 is used as the training set and Smoke-12 is as the test set. Firstly, data enhancement is performed on test set smoke-12. Classical and current mainstream recognition algorithms are used including EDL [32], DDCNN [33], EPHD [34]. In contrast, proposed GAN performs data enhancement [35] on training set Normal-12 and generates data similar to Smoke scenes for testing on Smoke-12.

To verify the segmentation effect of the method on smoke in different scenarios, the images of the test set are divided into five scenarios: easily distinguishable scene, the small-area smoke scene, the thin smoke scene, the multi-objective smoke scene, and the cloud interference scene, as shown in Figure 6.

Fig. 6. Scene classification scheme (from right to left: easily distinguishable scene, the small-area smoke scene, the thin smoke scene, the multi-objective smoke scene, and the cloud interference scene)

GAN is the result of using normal-12 as the training set and smoke-12 as the test set in Table 2. It can be seen from Table 2 that MCW, ALSTM, KutralNext algorithms have an increase of 0.8%, 1.5% and 1.9% in classification accuracy compared with GAN respectively. The DCP algorithm shows a 1.6% decrease. According to the experiment, it can be seen that it is difficult to achieve good results in the real diversified smoke scene by using the algorithm of defogging data enhancement. The recognition model developed from the synthetic smoke data set used by most recognition algorithms will over-fit on the synthetic data set and fail to narrow the distribution of the pixel field. Recognition algorithms will cause some information loss in the image. According to the results of proposed GAN in Table 2, the new GAN data enhancement can achieve the best results in three basic classification and recognition networks, and the accuracy is significantly improved compared with direct recognition of foggy images.

Table 2. Data enhancement with different methods/%

Method	GAN	MCW	ALSTM	KutralNext	Proposed
truck	58.5	53.9	53.3	55.0	65.5
tree	92.1	98.2	95.6	97.7	98.1
train	45.7	56.4	66.1	71.4	72.9
streetlamp	94.8	87.2	90.7	93.6	94.5
plane	81.4	79.7	88.8	81.9	91.8
car	51.8	54.7	49.8	41.2	62.5
bus	33.1	33.6	46.7	48.2	49.9
building	76.8	90.6	88.2	90.6	91.7
bridge	97.2	93.9	93.5	94.6	98.2

According to the experimental results in Table 3, the classification accuracy of CBAM module is improved by 3.3%. The dual channel can fully align the feature distribution in the feature domain and effectively reduce the field offset.

Figure 7 and table 4 shows the change curve of the proposed GAN loss in the industrial smoke data set. It can be seen from Figure 6 that the loss curve of the training set is ideal

Table 3. Classification accuracy after processing with proposed GAN/%

Method	GAN	EDL	DDCNN	EPHD	Proposed
truck	65.5	64.9	70.8	71.2	72.7
tree	94.1	95.6	95.8	96.2	97.3
train	72.4	78.2	81.6	80.2	96.1
streetlamp	94.2	88.9	89.5	91.3	95.5
plane	91.8	91.1	92.3	92.4	94.6
car	62.5	69.3	71.4	72.5	73.8
bus	54.6	46.5	53.8	58.1	62.2
building	91.6	91.8	92.3	92.7	94.1
bridge	98.2	99.1	97.2	98.5	99.1

and the network fitting speed is fast. In 10 epoches, the training is relatively stable, the network convergence speed is fast, and there is no fitting phenomenon.

Table 4. The training loss

epoch	0	10	20	30	40	50
loss	1.12	0.52	0.50	0.48	0.48	0.48

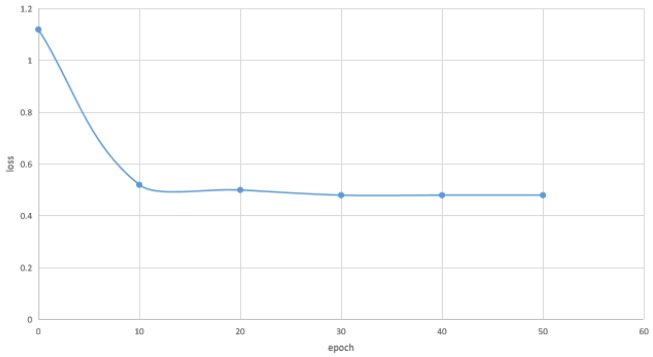

Fig. 7. The curve of training loss

In order to quantitatively analyze the recognition and segmentation results of industrial smoke, Precision, Recall, F-score, Accuracy [56] and Intersection over Union (IoU) [35] are selected as the evaluation indexes.

The same smoke data set and hyperparameter are used to train four models respectively, and then tested on the test sets with the five scenes. The comparison of recognition results is shown in figures 8-12. (a) is the original image, (b) is the result of manual marking, (c) is the EDL model, (d) is the result of DDCNN, (e) is the result of EPHD, and (f)

is the result of the Proposed method. The quantitative index pairs of test results are shown in Tables 5-9.

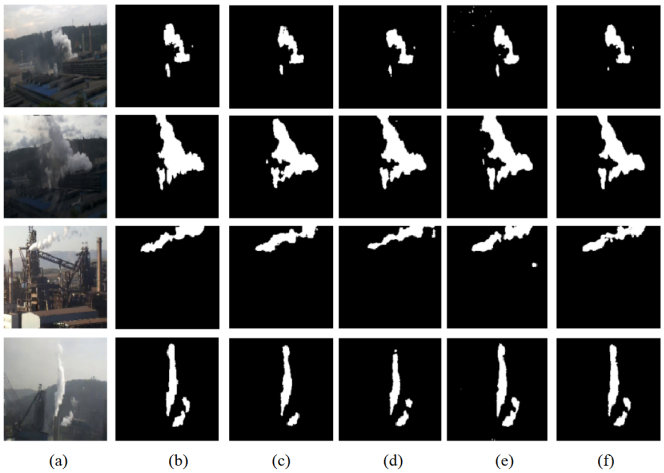

Fig. 8. Comparison on easily distinguishable scene

Table 5. Comparison result of easily distinguishable scene

Index	EDL	DDCNN	EPHD	Proposed
IoU/%	68.26	69.52	84.12	83.74
Precision/%	88.18	86.68	91.24	92.23
Recall/%	75.96	78.82	92.49	91.18
F1-score/%	81.61	84.98	91.86	91.70

6.1. Experimental Results and Discussion

1. In the comparison of distinguishable scene recognition results, interference points appear in the results of EDL, and a part of the background region is segmented. DDCNN has the problem of missing small areas, which leads to incomplete recognition results. There is a small amount of interference in EPHD results. The method presented in this paper is the most accurate.
2. In the comparison of small area smoke scene recognition results, a part of non-smoke area is segmented by EDL. The segmentation result of DDCNN is not accurate, the result is incomplete and the problem of false recognition occurs. EPHD also has some inaccurate results. The method in this paper is the most accurate.
3. In the comparison of thin smoke scene recognition results, EDL has the problem of identifying the background of small areas as smoke. DDCNN results are incomplete. The results of EPHD and the proposed method are relatively accurate.

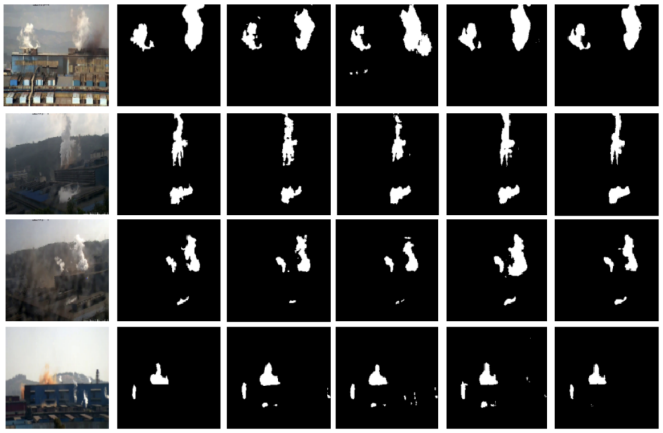

Fig. 9. Comparison on multi-objective smoke scene

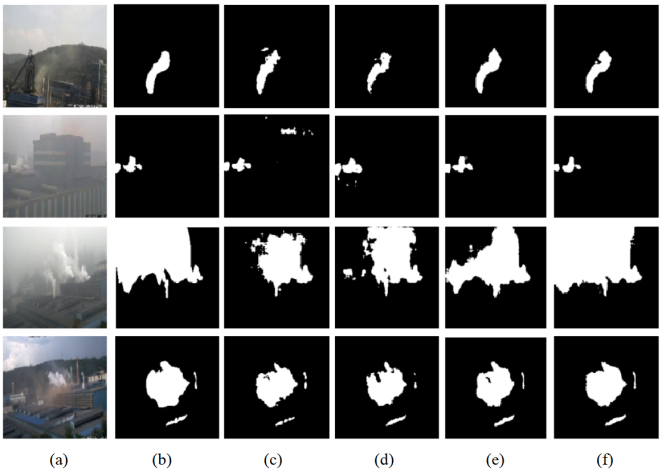

Fig. 10. Comparison on thin smoke scene

Table 6. Comparison result of small-area smoke scene

Index	EDL	DDCNN	EPHD	Proposed
IoU/%	58.22	60.48	77.22	78.86
Precision/%	77.24	81.76	83.55	89.28
Recall/%	68.22	69.09	92.28	87.30
F1-score/%	75.44	78.86	87.69	88.28

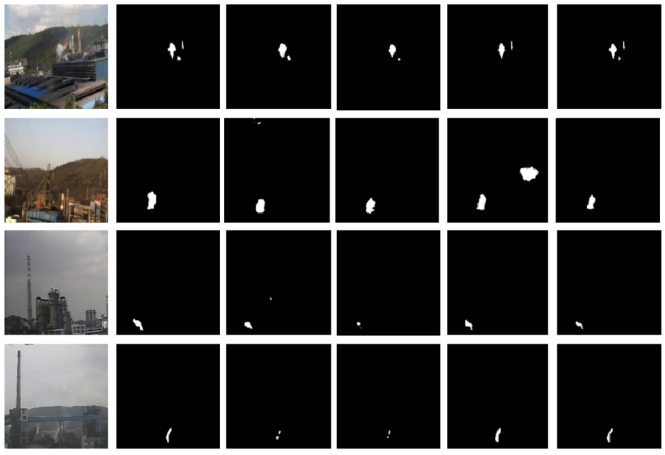

Fig. 11. Comparison on small-area smoke scene

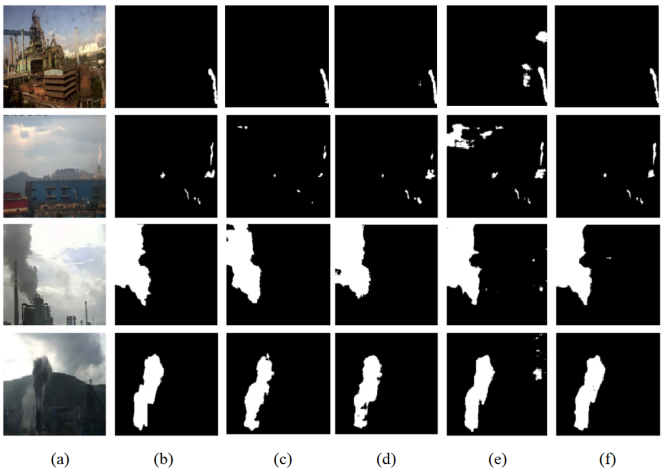

Fig. 12. Comparison on cloud interference scene

Table 7. Comparison result of multi-objective smoke scene

Index	EDL	DDCNN	EPHD	Proposed
IoU/%	72.04	70.21	79.66	80.05
Precision/%	88.61	86.94	87.80	90.19
Recall/%	81.14	80.44	90.93	88.89
F1-score/%	84.71	85.53	89.34	89.53

Table 8. Comparison result of cloud interference scene

Index	EDL	DDCNN	EPHD	Proposed
IoU/%	65.19	66.58	72.69	78.27
Precision/%	82.56	83.27	79.55	88.74
Recall/%	77.66	79.09	90.16	87.50
F1-score/%	80.04	82.40	84.52	88.12

Table 9. Comparison result of cloud interference scene

Index	EDL	DDCNN	EPHD	Proposed
IoU/%	69.21	67.85	76.41	75.75
Precision/%	85.94	85.52	84.79	89.29
Recall/%	79.97	78.62	89.71	89.98
F1-score/%	82.85	84.04	87.18	88.08

4. In the comparison of multi-object smoke scene recognition results, both EDL and DDCNN have the problem of excessive segmentation, and the edge of recognition results is not accurate. EPHD divides a small building area. EPHD fails to identify the darker soot. The new method has the best relative performance.
5. In the comparison of recognition results of interference scenes, EDL is greatly affected by cloud interference, and recognition results include areas belonging to clouds. DDCNN and EPHD have better anti-interference than EDL, but the results are not complete. The new method shows the best anti-interference performance.

To sum up, the anti-interference ability of EDL is poor, and it is easy to identify and segment the non-smoke area. Moreover, there is the problem of excessive segmentation, which is reflected in the inaccurate edge of the identification result. With the encoding and decoding structure adopted by DDCNN, the precision of the recognition results should be better than that of EDL. However, because the smoke edges are rough and the image resolution used in the experiment is not high, the edge improvement is not obvious from the results, but the anti-interference ability is obviously better than that of EDL, but the recognition results are incomplete. On the basis of DDCNN, EPHD adds a parallel shallow network to refine the results. As can be seen from the recognition results, EPHD has a certain degree of improvement in anti-interference compared with DDCNN, and also has a great improvement in the problem of missing small areas that DDCNN is prone to. The proposed method has the best anti-interference performance among all the methods. The multi-scale convolution operation adopted enhances the feature extraction ability of smoke. Compared with EDL, the test results can effectively distinguish the interference in soot and background, and the comparison results are closest to the manual labeling results.

In terms of quantitative indicators, DDCNN method has the lowest index, because the pre-trained network weight is not used in the initialization of the network parameters. In order to improve the accuracy of recognition results, a parallel branch is added. To compare the results of the new parallel network pair. During the training, it maintains the same parameter initialization mode as DDCNN. In terms of indicators, it has a certain improvement compared with DDCNN. The method adopted in this paper has significantly

improved the accuracy rate and intersection ratio index of test results in all scenarios compared with EDL, especially in the test set of interference scenes, which is also consistent with the performance of recognition results comparison.

7. Conclusion

Considering that it is difficult to achieve good classification and recognition effect for unlabeled smoke scene images under existing labeled data, this paper proposes a GAN model based on LSTM and CBAM. The existing data is enhanced to make it easily and effectively close to the smoke scene in the style and data distribution of the pixel field. And LSTM module is used to make full use of data feature information through multi-scale feature module to further align data distribution in feature domain effectively. Combined with CBAM, the accuracy of smoke classification and identification is greatly improved. Through experiments, the proposed framework has significantly improved the classification and recognition accuracy in smoke scenes compared with the basic classification network trained with existing labeled data. The current method still needs to be further improved in classification accuracy and structural optimization. The subsequent research will explore the fusion of algorithms in pixel domain and feature domain, improve the overall framework structure, further enrich the data set, and make the overall performance towards a more efficient direction.

Acknowledgments. This work was supported by Henan Province Science and Technology Project (212102210273).

References

1. Niero M, Ingvordsen C H, Peltonen-Sainio P, et al. "Eco-efficient production of spring barley in a changed climate: A Life Cycle Assessment including primary data from future climate scenarios," *Agricultural Systems,* vol. 136, pp. 46-60. (2015)
2. Arora N K. Impact of climate change on agriculture production and its sustainable solutions[J]. Environmental Sustainability, 2019, 2(2): 95-96.
3. Kumar V S, Muthukumaravel A. Seasonal forecasting of mobile data traffic in GSM networks with linear trend[J]. Journal of Applied Science and Engineering, 2020, 23(3): 469-474.
4. Ousmen A, Touraine C, Deliu N, et al. Distribution-and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review[J]. Health and quality of life outcomes, 2018, 16(1): 1-12.
5. Chen W T, Ding J J, Kuo S Y. PMS-net: Robust haze removal based on patch map for single images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11681-11689.
6. Zhang H, Patel V M. Densely connected pyramid dehazing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3194-3203.
7. Huang S, Liu Y, Wang Y, et al. A new haze removal algorithm for single urban remote sensing image[J]. IEEE Access, 2020, 8: 100870-100889.
8. Zhu Q, Du B, Yan P. Boundary-weighted domain adaptive neural network for prostate MR image segmentation[J]. IEEE transactions on medical imaging, 2019, 39(3): 753-763.
9. Chen C, Chen Z, Jiang B, et al. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 3296-3303.

10. Liguo Wang, Yin Shoulin, Hashem Alyami, et al. A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images [J]. Geoscience Data Journal, 2022. https://doi.org/10.1002/gdj3.162
11. Man Jiang and Shoulin Yin. Facial expression recognition based on convolutional block attention module and multi-feature fusion [J]. Int. J. of Computational Vision and Robotics, 2021. DOI:10.1504/IJCVR.2022.10044018
12. Gallego A J, Calvo-Zaragoza J, Fisher R B. Incremental unsupervised domain-adversarial training of neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(11): 4864-4878.
13. Zhao Y, Wu R, Dong H. Unpaired image-to-image translation using adversarial consistency loss[C]//European Conference on Computer Vision. Springer, Cham, 2020: 800-815.
14. Chen J, Chen L, Wang S, et al. A novel multi-scale adversarial networks for precise segmentation of x-ray breast mass[J]. IEEE Access, 2020, 8: 103772-103781.
15. Karnewar A, Wang O. Msg-gan: Multi-scale gradients for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7799-7808.
16. Zhang, J., Yu, X., Lei, X., Wu, C.: A Novel Deep LeNet-5 Convolutional Neural Network Model for Image Recognition. Computer Science and Information Systems, Vol. 19, No. 3, 1463-1480. (2022), https://doi.org/10.2298/CSIS220120036Z.
17. Sonekar S V, Pal M, Tote M, et al. Enhanced route optimization technique and design of threshold-T for malicious node detection in ad hoc networks[J]. International Journal of Information Technology, 2021, 13(3): 857-863.
18. de Oliveira Khn V, Lopes B C F L, Caicedo B, et al. Micro-structural and volumetric behaviour of bimodal artificial soils with aggregates[J]. Engineering Geology, 2021, 288: 106139.
19. Teng, L., Qiao, Y.: BiSeNet-oriented context attention model for image semantic segmentation. Computer Science and Information Systems, Vol. 19, No. 3, pp. 1409-1426. (2022), https://doi.org/10.2298/CSIS220321040T
20. Wu Z, Xue R, Li H. Real-Time Video Fire Detection via Modified YOLOv5 Network Model[J]. Fire Technology, 2022, 58(4): 2377-2403.
21. Li Y, Wen W, Guo X, et al. High-throughput phenotyping analysis of maize at the seedling stage using end-to-end segmentation network[J]. PLoS One, 2021, 16(1): e0241528.
22. Mazzia V, Angarano S, Salvetti F, et al. Action Transformer: A self-attention model for short-time pose-based human action recognition[J]. Pattern Recognition, 2022, 124: 108487.
23. Li Y, Ko Y, Lee W. RGB image-based hybrid model for automatic prediction of flashover in compartment fires[J]. Fire safety journal, 2022, 132: 103629.
24. Smagulova K, James A P. A survey on LSTM memristive neural network architectures and applications[J]. The European Physical Journal Special Topics, 2019, 228(10): 2313-2324.
25. Gui J, Sun Z, Wen Y, et al. A review on generative adversarial networks: Algorithms, theory, and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.
26. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
27. Yang F, Moayedi H, Mosavi A. Predicting the degree of dissolved oxygen using three types of multi-layer perceptron-based artificial neural networks[J]. Sustainability, 2021, 13(17): 9898.
28. Ju A, Wang Z. A novel fully convolutional network based on marker-controlled watershed segmentation algorithm for industrial soot robot target segmentation[J]. Evolutionary Intelligence, 2022: 1-18.
29. Cao Y, Yang F, Tang Q, et al. An attention enhanced bidirectional LSTM for early forest fire smoke recognition[J]. IEEE Access, 2019, 7: 154732-154742.
30. Ayala A, Macdo D, Zanchettin C, et al. KutralNext: An Efficient Multi-label Fire and Smoke Image Recognition Model[C]//Anais Estendidos do XXXIV Conference on Graphics, Patterns and Images. SBC, 2021: 7-13.

31. Yang Z, Wang T, Bu L, et al. Training with augmented data: Gan-based flame-burning image synthesis for fire segmentation in warehouse[J]. Fire Technology, 2022, 58(1): 183-215.
32. Namozov A, Im Cho Y. An efficient deep learning algorithm for fire and smoke detection with limited data[J]. Advances in Electrical and Computer Engineering, 2018, 18(4): 121-128.
33. Gu K, Xia Z, Qiao J, et al. Deep dual-channel neural network for image-based smoke detection[J]. IEEE Transactions on Multimedia, 2019, 22(2): 311-323.
34. Yuan F, Shi J, Xia X, et al. Encoding pairwise Hamming distances of Local Binary Patterns for visual smoke recognition[J]. Computer Vision and Image Understanding, 2019, 178: 43-53.
35. Y. Yuan, Z. Xu and G. Lu, "SPEDCCNN: Spatial Pyramid-Oriented Encoder-Decoder Cascade Convolution Neural Network for Crop Disease Leaf Segmentation," in IEEE Access, vol. 9, pp. 14849-14866, 2021, doi: 10.1109/ACCESS.2021.3052769.

Dahai Li is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction: Information security, image processing.

Yang Rui is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064, China. Research direction: Information security, image processing.

Su Chen is with the Department of Mechanical and Electrical Engineering, Henan Vocational College of Water Conservancy and Environment, Zhengzhou 450002, China. Research direction: image processing.

Received: November 25, 2022; Accepted: March 30, 2023.