photonics

MDPI

Article

A Fast Accurate Attention-Enhanced ResNet Model for Fiber-Optic Distributed Acoustic Sensor (DAS) Signal Recognition in Complicated Urban Environments

Xinyu Liu [1], Huijuan Wu [1,*], Yufeng Wang [1], Yunlin Tu [1], Yuwen Sun [1], Liang Liu [1], Yuanfeng Song [1], Yu Wu [1] and Guofeng Yan [2]

[1] Key Laboratory of Optical Fiber Sensing and Communications (Ministry of Education), School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2] Fiber Optic Sensing Research Center, Zhejiang Laboratory, Hangzhou 310000, China
* Correspondence: hjwu@uestc.edu.cn; Tel.: +86-136-9901-0975

Abstract: The fiber-optic distributed acoustic sensor (DAS), which utilizes existing communication cables as its sensing media, plays an important role in urban infrastructure monitoring and natural disaster prediction. In the face of a wide, dynamic environment in urban areas, a fast, accurate DAS signal recognition method is proposed with an end-to-end attention-enhanced ResNet model. In preprocessing, an objective evaluation method is used to compare the distinguishability of different input features with the Euclidean distance between the posterior probabilities classified correctly and incorrectly; then, an end-to-end ResNet is optimized with the chosen time-frequency feature as input, and a convolutional block attention module (CBAM) is added, which can quickly focus on key information from different channels and specific signal structures and improves the system recognition performance further. The results show that the proposed ResNet+CBAM model has the best performance in recognition accuracy, convergence rate, generalization capability, and computational efficiency compared with 1-D CNN, 2-D CNN, ResNet, and 2-D CNN+CBAM. An average accuracy of above 99.014% can be achieved in field testing; while dealing with multi-scenario scenes and inconsistent laying or burying environments, it can still be kept above 91.08%. The time cost is only 3.3 ms for each signal sample, which is quite applicable in online long-distance distributed monitoring applications.

Keywords: DAS recognition; Φ-OTDR; time-frequency characteristics; ResNet; attention mechanism

Citation: Liu, X.; Wu, H.; Wang, Y.; Tu, Y.; Sun, Y.; Liu, L.; Song, Y.; Wu, Y.; Yan, G. A Fast Accurate Attention-Enhanced ResNet Model for Fiber-Optic Distributed Acoustic Sensor (DAS) Signal Recognition in Complicated Urban Environments. *Photonics* **2022**, *9*, 677. https://doi.org/10.3390/photonics9100677

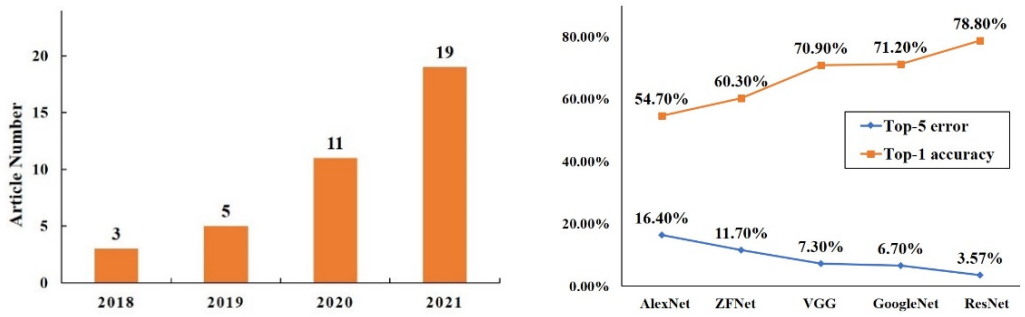
Received: 24 August 2022
Accepted: 16 September 2022
Published: 21 September 2022

1. Introduction

The distributed optical fiber acoustic sensor (DAS), based on phase-sensitive optical time-domain reflectometry (Φ-OTDR) technology [1,2], provides a highly sensitive and cost-effective continuous sensing method for dynamic acoustics or vibrations in long distances and wide ranges. It has been extensively applied in urban infrastructure monitoring [3–5], natural disaster prediction [6] applications, energy exploration in the oil and gas industry [7], etc. Researchers have made great efforts to improve its hardware index [8–10] and detection and identification abilities [11–18]. Furthermore, lots of useful feature extraction and classifier design work [19–28] has been involved. The perception and recognition ability of Φ-OTDR is, thus, enhanced, but these methods rely heavily on expert knowledge to build their systems. With the rapid development of Artificial Intelligence (AI) technology, more and more deep learning techniques are being applied to DAS, as shown in Figure 1a. In particular, the convolutional neural network (CNN) [29–36], bilinear-CNN (B-CNN) [37], and one-dimensional CNN (1-D CNN) [33,38], as well as combined models such as improved multi-scale-CNN with hidden-Markov-model (mCNN-HMM) [39], attention-enhanced long short-term memory (ALSTM) [40], and squeeze and excitation

WaveNet (SE-WaveNet) [41] have been successively used, which means it is more and more convenient for us to extract the hidden features of different signal targets automatically, and this also obtains improved results.

Among these networks, CNN is a typical representative due to its outstanding local feature extraction ability with its convolution blocks [42]. However, it also has some common problems, such as gradient disappearance/explosion and local convergence in the training of the deep network. Thus, various CNN network structures have been developed, and a performance comparison based on the ImageNet data set is shown in Figure 1b. For example, the deep residual network, ResNet, was proposed by He et al. [43] in 2015 to eliminate redundant layers autonomously and greatly solve the problem of gradient disappearance. In addition, it has the best performance, as shown from its top-5 error and top-1 accuracy (top-1 accuracy is the probability that the number of samples was correctly labeled in the maximum probability of the model output over the total number of samples; top-5 error is the probability that all five outcomes of the assumed model-predicted outputs are wrong). Simultaneously, the importance of the attention mechanism has also been noticed by researchers [44–46]. In 2019, Chen et al. [40] proposed a long short-term memory (LSTM) model with an attention mechanism called ALSTM for signal recognition in DAS. However, attention is realized by a weighted sum of different features, which cannot be updated automatically. Moreover, the feature extraction of ALSTM is time-consuming and laborious, resulting in a poor real-time performance. At present, it is still challenging to find an accurate DAS recognition method with good generalization and high computation efficiency in the face of a wide, dynamic environment in urban areas.



(**a**) (**b**)

Figure 1. (**a**) Deep learning in DAS from 2018 to 2021; (**b**) performance comparison of different CNN architectures based on ImageNet data set [47].

Therefore, a fast, accurate, attention-enhanced ResNet model for DAS signal recognition is proposed in this paper. The main contribution includes:

(1) An end-to-end ResNet network is proposed for DAS signal recognition. The recognition accuracy can be further improved by 1.3%, and the time for each signal sample can be saved by 40% compared to the common 2-D CNN network. This also shows the training and the online test processes both speed up through the residual blocks in ResNet. Furthermore, the generalization capability is also improved a lot in more challenging, atypical, and inconsistent signal recognition under multi-scenario conditions.

(2) An attention module with a CBAM is added to the ResNet network, which enables the model to focus on the key features of the signal quickly and automatically through both the local structure and channel attention. This highlights the difference between significant structural information and channel information, thus achieving a high recognition rate of 99.014% for four typical DAS events in urban communication cable monitoring.

(3) The effectiveness of different methods of extracting deep features is evaluated via the Euclidean distance between the posterior probabilities, classified correctly and incorrectly, which are calculated in a matrix. It assumes when the Euclidean distances between the posterior probabilities, classified correctly and incorrectly, for one type

of sample are more considerable, the degree of feature distinguishability is stronger. In this way, different models' feature extraction capabilities can be measured by an objective parameter rather than only based on classification accuracy.

2. Recognition Methodology with the Attention-Enhanced ResNet Model in DAS

2.1. Data Collection with DAS

The typical structure of DAS and its working principle in the safety monitoring of communication cables buried in urban areas are shown in Figure 2. It usually directly takes in a dark fiber from the communication cable laid under urban ground. The system hardware consists of three parts: the detection cable, the optical signal demodulation equipment, and the signal processing host. The optical signal demodulation equipment mainly includes optical devices and electrical devices. A continuous, coherent optical signal is generated by an ultra-narrow linewidth laser, and the optical pulse signal is modulated by an audio-optical or electro-optical modulator. The optical pulse signal is concentrated and amplified by an erbium-doped fiber amplifier (EDFA), and the amplified optical pulse signal is injected into the detection cable by an isolator and a circulator in turn. A light pulse signal along the fiber optic cable transmission process produces Rayleigh scattering; subsequent to the Rayleigh scattering, a light signal along the cable is returned and received by the circulator. Then, the phase change in the coherent Rayleigh backscattering light wave carrying the vibration information is linearly demodulated by an imbalanced Mach–Zehnder fiber interferometer (MZI) and a 3 × 3 coupler [48], as shown in Figure 2. Each point in the fiber is equivalent to a sensor node. These distributed nodes cooperate to sense vibration signals in the whole line. Thus, the system returns a space–time matrix as:

$$\{XX = x_{ts}\ (t = 1,\ 2, \cdots, T; s = 1,\ 2, \cdots, S)\} \tag{1}$$

where the row index, t, represents the time; T is the time length; the column index, s, denotes the spatial sampling node; and S is the spatial width. The spatial interval of each two nodes is ΔS, and the temporal interval is $\Delta T = \frac{1}{f_s}$, in which f_s is the sampling frequency. One data column represents the temporal signal collected at a sampling node.

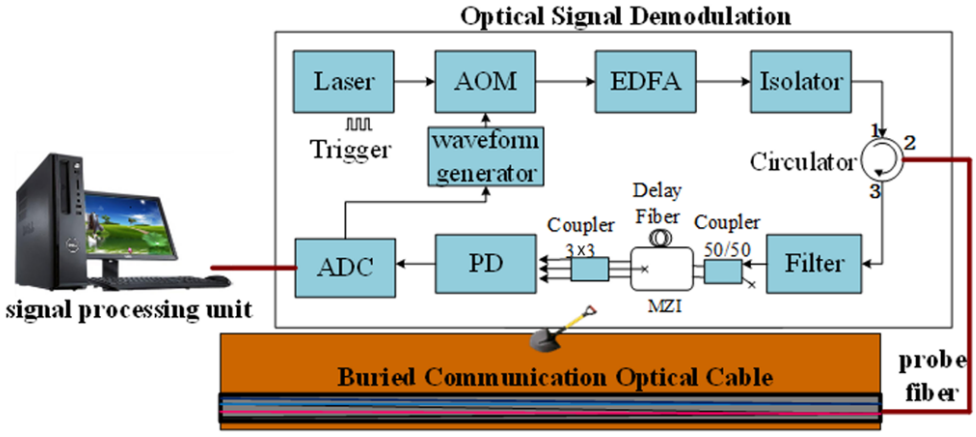


Figure 2. Experimental setup for data collection in the field.

In the cable monitoring task, four frequently encountered events are included: background noises, traffic interference, manual digging, and mechanical excavations, which are labeled as 0 to 3, respectively. In the recognition, a 1-D temporal signal with a certain length at the event's location is selected from the time–space matrix and built into the data sets. The flowchart of the proposed attention-enhanced ResNet model is illustrated in Figure 3, which includes three stages: data preprocessing, the construction of the attention-enhanced ResNet network, and offline training and online testing of the network.

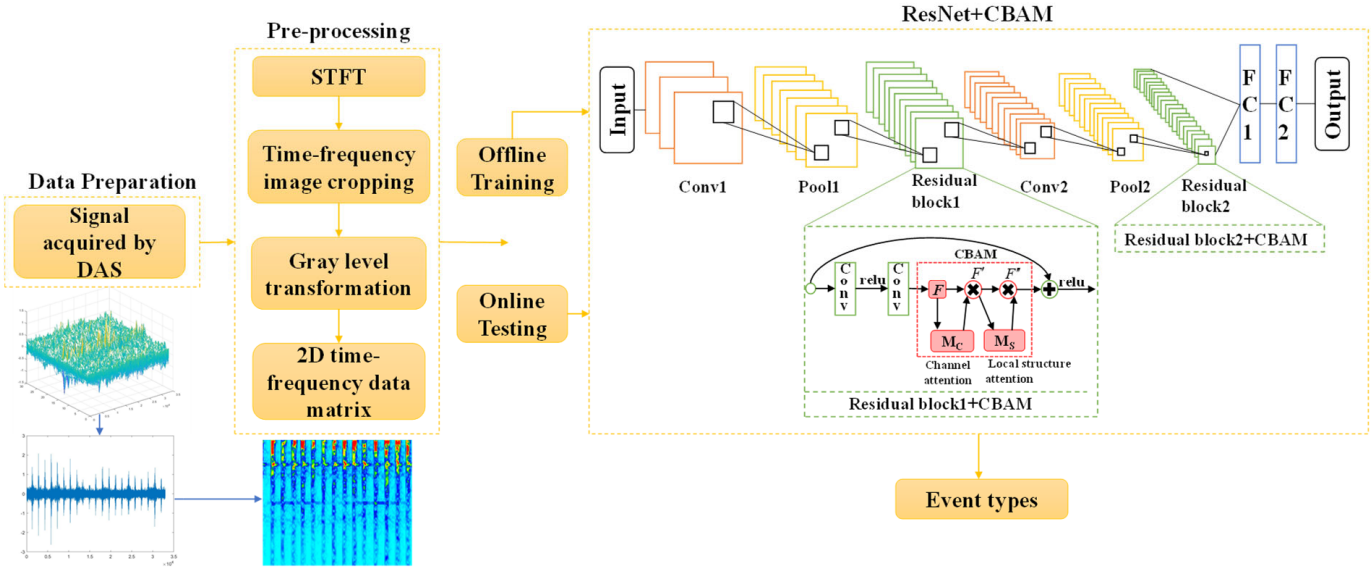


Figure 3. Flowchart of the proposed attention-enhanced ResNet model in DAS.

2.2. Data Preprocessing

In the data preprocessing, each signal sample is converted into a time-frequency spectrogram with short-time Fourier transform (STFT). In detail, the STFT is realized on the DAS signal, $x(n)$:

$$X_n(\omega) = \sum_{n=mR}^{n=mR+L-1} x(n)w(n-mR)e^{-jwn} \quad (2)$$

where $w(n)$ is a rectangular window of length L used to obtain the windowed data frame for the STFT, R is the hop size of the window, and mR $(m = 1, 2, 3, \ldots)$ is the moving location of the windowed data, as the window "slides" or "hops" over time. The window length, L; the hop size, R; and the FFT points, $nFFT$, are three critical parameters that need to be carefully chosen in applications, or they will influence the quality of the spectrogram and its time consumption.

When the database is ready, the processing of the STFT, the image preprocessing of the gray conversion, and clipping for the spectrogram are successively carried out. In this application, the sampling rate is 500 Hz, and the sample length is 10 s. To ensure the resolution of the spectrogram in the STFT, a boxcar window with a 95 data length (about 0.2 s) is chosen, the hop size is one sample (0.2 ms), and the FFT size is equal to the window length. The time-frequency matrix is built in a linear way without the logarithmic operation. To alleviate the computational load of the following network, the obtained time-frequency spectrogram is converted into a gray image with gray levels of 0–255, clipped, and then resized by downsampling it into a smaller size of 50 (in frequency axis) $\times$ 100 (in time axis). In total, 50 stands are used for a clipped frequency range of 125 Hz and 100 are used for the 10 s sample length time range. The purpose of clipping is to reduce the image dimension; then, a 2-D time-frequency 50 $\times$ 100 data matrix is obtained as the input of the following ResNet network.

2.3. Attention-Enhanced ResNet Network Construction

To adapt to the time-frequency characteristics of the DAS signals, the basic residual block in ResNet is composed of convolution layers and rectified linear unit (ReLU) layers. Assuming the input of the *l-th* residual block is x_l, its output is formalized in one of the following two ways:

If x_l and $F(x_l, W_l)$ have the same dimension

$$x_{l+1} = f(x_l + F(x_l, W_l)) \quad (3)$$

then

$$x_{l+1} = f(W_s x_l + F(x_l, W_l)) \tag{4}$$

where $F(x_l, W_l)$ is the residual function; W_l is the weight parameter of $F(x_l, W_l)$; $f(\cdot)$ is the ReLU; and W_s is a linear mapping, which can be performed through a shortcut connection to match their dimensions.

Further, in the residual block, a convolutional block attention module (CBAM) [49] is added, as shown in Figure 3, which is selected from Table 1 to secure the highest recognition performance and training efficiency. CBAM deduces the attention map from the intermediate feature map, F, along the channel and the local structure dimensions in turn. Specifically, the attention map is multiplied by the input feature map for adaptive feature refinement, emphasizing the meaningful features along the two main dimensions of channel and local structure in the convolution process. Then, the calculation process of the final feature map F'' is:

$$F' = M_C(F) \otimes F \tag{5}$$

$$F'' = M_S(F') \otimes F' \tag{6}$$

where F' is the feature map output after the channel attention module; $\otimes$ represents element-wise multiplication; M_C is the channel attention map; and M_S is the local structure attention map. M_C and M_S are calculated specifically as follows:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{7}$$

$$M_S(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \tag{8}$$

where σ represents the Sigmoid function; MLP represents the multilayer perceptron; $f^{7\times7}$ represents a convolution operation; and the filter size is 7×7.

Table 1. Performance comparison of different attention modules [50].

Attention Modules	Recognition Accuracy	Training Time/s
None	84.52%	11
SE-NET	92.21%	11
ECA-NET	92.28%	11
SK-NET	90.66%	13
CBAM	**93.23%**	**12**
DANET	88.37%	12
PFAN	92.47%	12

2.4. Training of the Network

Similar to CNN, ResNet+CBAM is generally trained through feedforward, backpropagation, and iterative parameter updating.

Initialization: Typical parameters of weight, W, and bias, b, are initialized using the Xavier initialization method, and the uniform distribution range is chosen as:

$$\begin{cases} W \sim U[-\frac{\sqrt{6}}{\sqrt{n_{in} + n_{out}}}, \frac{\sqrt{6}}{\sqrt{n_{in} + n_{out}}}] \\ b \sim U[-\frac{\sqrt{6}}{\sqrt{n_{in} + n_{out}}}, \frac{\sqrt{6}}{\sqrt{n_{in} + n_{out}}}] \end{cases} \tag{9}$$

where n_{in} is the number of input parameters and n_{out} is the output parameter number.

Convolution layer: The convolution is calculated as:

$$conv_u = \sum_{i,j} X_i W_j + b, u \in (1, floor(\frac{L - m + 2p}{s})) \tag{10}$$

where X_i is the *i-th* input, W_j is the weight matrix after initialization of the *j-th* convolution kernel, m is the kernel size of the convolution, s is the stride, p is the padding, and L is the input sequence length. Then, the convolution output is activated by ReLU.

Pooling layer: The pooling is calculated as:

$$avg\_pool = \left\{ avg\left[conv_u^1, conv_u^2, \ldots, con \ldots\ldots avg[conv_u^p, conv_u^{\cdots 1}, \ldots, conv_u^{p+s}] \right\} \quad (11)$$

where s is the stride and p is the padding.

Fully connected layer: The final classification output, y, is obtained as:

$$y = W \times x + b \quad (12)$$

where x is the input, W is the weight matrix, and b is the bias. Specifically, a dropout layer is added to avoid the overfitting phenomenon.

Loss function: Cross entropy is used as the loss function to train the whole ResNet+CBAM network:

$$L = -\frac{1}{N}[y \ln a + (1 - y) \ln(1 - a)] \quad (13)$$

where N is the batch size or the sample number in the data batch for training; y is the true label of the sample; and a is the predicted label of the sample.

3. Real Field Test Results and Discussion

3.1. Field Data Collection and Preprocessing

The data were collected by the DAS system, as shown in Figure 2, to monitor the underground communication cables in two cities in China: Tongren in Guizhou Province and Wuhan in Hubei Province. We mixed the event data from the two cities: one 34 km underground cable located in Wuhan, Hubei Province and the other 18 km underground cable located in Tongren, Guizhou Province. The key data collection parameters for the two monitoring fields were the same. The selected cables were buried 0.8–1.5 m deep in the underground; the linewidth of the laser was 5 kHz; the pulse width was 200 ns, which corresponds to a spatial resolution (gauge length) of 20 m; the temporal sampling rate of the system was 500 Hz; and its spatial sampling interval was 5.16 m. These real event data were collected at different locations of the whole line. However, some mechanical excavation and manual digging were present at two or three selected locations along each line. The data were collected in a period, but not in a day. Our test in Tongren lasted from February 2019 to March 2019, while the test in Wuhan lasted from April 2019 to May 2019. During the process of these months, the database obtained data from different weather conditions, different times, and different optical fiber locations. The diversity of the data was guaranteed so that the database has a good generalization ability to support our results and conclusions.

Here, four typical events, as stated above are chosen as our vibration targets to be identified, and the original signals and STFT time-frequency diagram obtained for the above four events are illustrated in Figure 4.

In the preprocessing, by using the training set in Table 2, STFT and Mel-Frequency Cepstral Coefficients (MFCC) methods are used to mine the time-frequency characteristics of the signals, and the Euclidean distance method is used to compare the resolution of the two time-frequency characteristics. MFCC has a higher resolution for the lower-frequency part, which can help to distinguish the changing background, traffic, manual digging, and mechanical excavation, which are all concentrated in the low-frequency band of 0–150 Hz. In detail, the basic steps of MFCC include Mel frequency conversion and cepstral analysis. The most critical steps of MFCC are as follows: 26 triangular filter banks are used to filter the power spectrum estimation, and then a logarithm was taken to perform DCT transform

to obtain the Mel-frequency Cepstral. The 12 retention numbers, from 2 to 13, are retained to obtain the MFCC features.

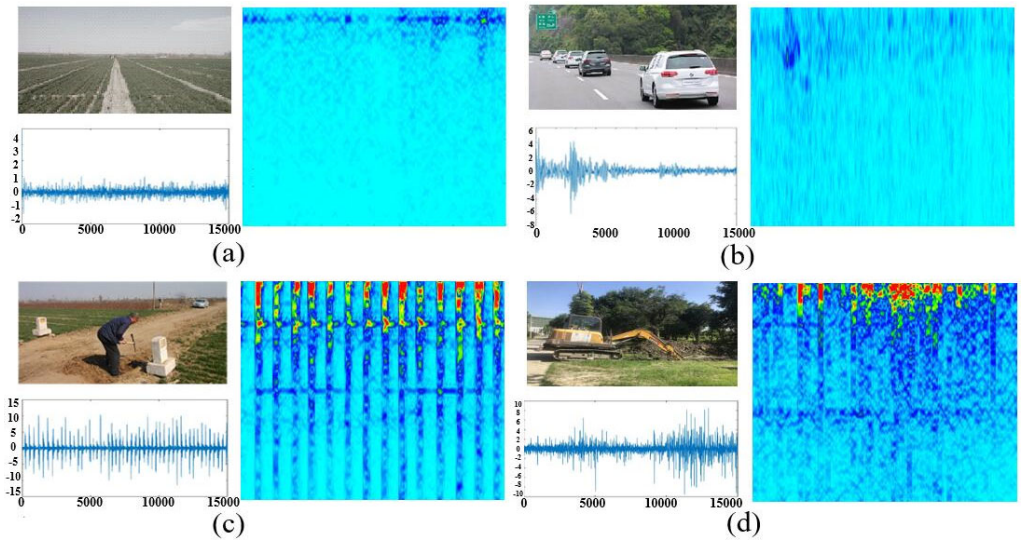


Figure 4. Field scenes, typical signals, and STFT pseudo-color RGB time-frequency diagrams of (**a**) background noise; (**b**) traffic interferences; (**c**) manual diggings; and (**d**) mechanical excavations.

Table 2. The database of the buried urban communication cable monitoring.

Time-Frequency Data Type	Training Size	Validation Size	Test Size	Event Label
Background noises	2730	910	910	0
Traffic interference	1756	585	585	1
Manual digging	780	260	260	2
Mechanical excavations	819	273	273	3

The distinguishability of the two types of time-frequency features is compared using the Euclidean distance index, which is defined as the distances between the posterior probabilities classified correctly and incorrectly for all the test samples, which is calculated as:

$$d_{im} = \frac{1}{N_{test}} \sqrt{\sum_{j}^{N_{test}} \left(p_i^j - p_m^j\right)^2}, \begin{cases} i, m = 0, 1, 2, 3 \\ m \neq i \end{cases}. \tag{14}$$

where N_{test} is the test sample number; j is the sample index; i and m represent the true and wrong labels, respectively; p_i^j is the posterior probability when it is correctly identified; and p_m^j is the posterior probability when it is incorrectly classified. In this way, the distances of the different posterior probability curves can be measured objectively. This assumes that when the distance between the posterior probabilities classified correctly and incorrectly for one type of sample is larger, the feature distinguishability is stronger.

According to Equation (14), the average Euclidean distances of STFT and MFCC can be calculated, and the time cost of time-frequency conversion for a single sample can also be calculated, as shown in Figure 5. It can be seen that the average distance of the STFT features is larger, indicating that STFT features have better distinguishability than MFCC features. Meanwhile, the time cost of the feature extraction of MFCC is three times longer than that of STFT. As a result, STFT is chosen for time-frequency feature extraction in this application, and the constructed database is detailed in Table 2. Each sample lasted 5 s. The training, validation, and test data are divided randomly according to a normal ratio of 6:2:2, and there is no duplication between them.

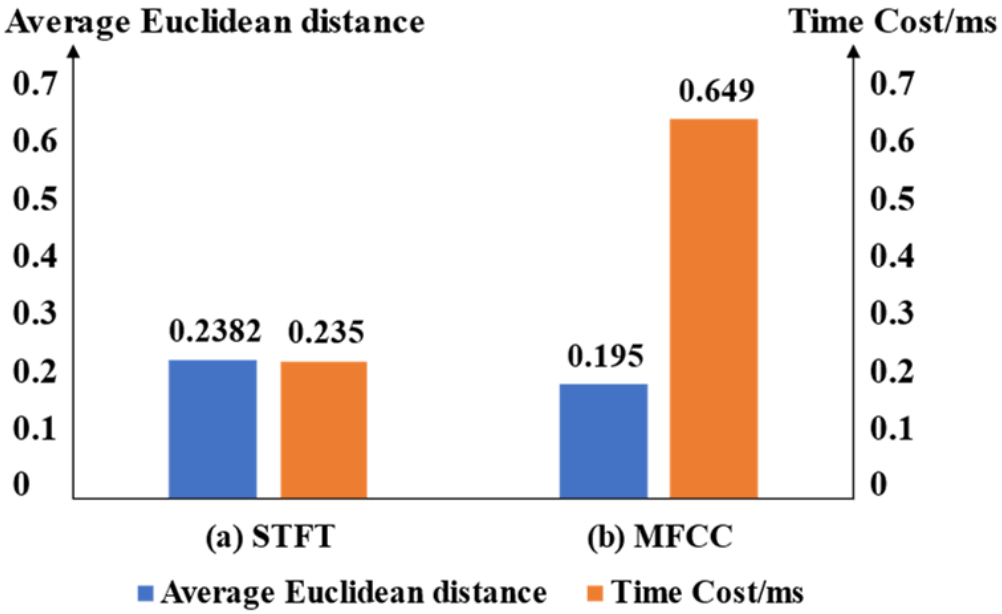


Figure 5. Comparison of the average Euclidean distances and the time cost with different features: (**a**) STFT, (**b**) MFCC.

3.2. Realization and Optimization of the Proposed ResNet+CBAM Network

In this section, we use the training set and validation set in Table 2 to realize and optimize the proposed network model. For the residual network (ResNet), there are some commonly used network structures such as ResNet18, ResNet34, ResNet50, ResNet101, and so on, in which the number represents the layer numbers of the network. The DAS signal recognition is different from the image recognition of multi-classification (>10) and multi-scene, whose data scene is relatively simple but requires higher real-time performance. Therefore, we do not consider the deep network and choose ResNet18, with its good real-time performance, as the basic architecture.

However, in the DAS signal recognition, to choose a proper number of residual blocks, different numbers of residual blocks for the training and validation are compared in Figure 6a,b with the same parameters. The training epoch is set to 10. The error bars for the means and standard deviations (3σ) of the training time and validation accuracy are obtained. It can be seen that, with the increase in the number of residual blocks, the training time and the validation accuracy both keep increasing, while the increasing rate of the validation accuracy decreases when the number exceeds two. Thus, in the proposed model in Figure 3, two residual blocks are used.

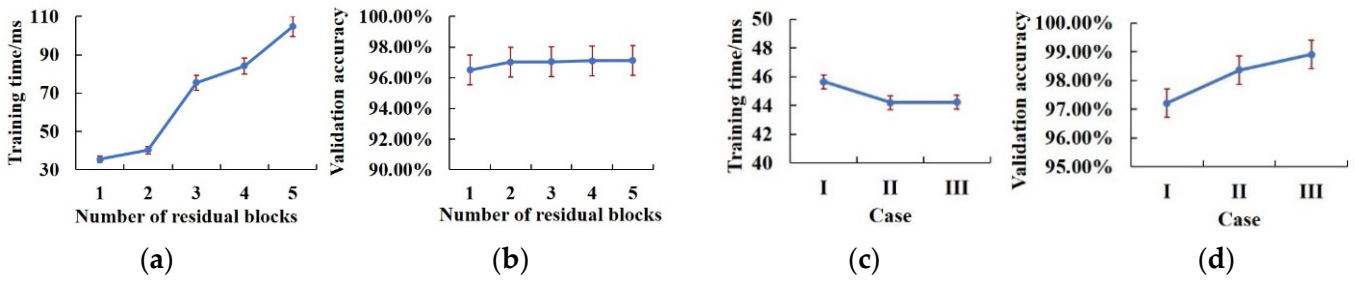


Figure 6. Validation results with different numbers of residual blocks, (**a**) training time, (**b**) accuracy; and validation results of the three cases, (**c**) training time, (**d**) accuracy.

As shown in Figure 3, there are two convolution layers and two residual blocks, among which there are two convolution layers in the residual block. Considering the definition of the CBAM, it can only work if it is added to the convolution layer of the network, and we

should find the optimal location of the CBAM. Thus, three possible positions of CBAM are compared: in Case I, the CBAM is outside the residual block and after Conv1 and Conv2, respectively; in Case II, it is located after the first Conv in the two residual blocks; and in Case III, it is located after the second Conv in the two residual blocks. These three cases can be roughly divided into two categories: adding a CBAM to the convolution layer outside (Case I) and inside of the two residual blocks (Cases II and III). The mean and standard deviation (3σ) error bars for the visualization total training time and the validation accuracy for the three network structures are shown in Figure 6c,d. This shows that the effect of the attention module inside the residual block is better than outside the residual block; it has higher validation accuracy and less training time cost. Further, the validation accuracy of 98.90% is higher in Case III than in Case II, which means the two residual blocks may play different roles, in which the first residual block possibly mines more common features, while the second one mines more specific features.

Meanwhile, in Figure 6, the training time of ResNet without a CBAM is 40.29 ms, while ResNet with a CBAM takes about 44.22 ms, indicating that the added CBAM takes less than 4 ms. It means the CBAM is a lightweight module that has little influence on the time cost of the whole network. Thus, the structure of Case III is selected.

In addition, the reduction number is a hyperparameter in the CBAM, which affects the ratio of input and output channels in the CBAM and needs to be carefully chosen through experimentation. In this paper, 16, 8, 4, 2, and 1 are tested and compared in the validation set. The default is 16, which is suitable for data sets with large sizes and large sample sizes. Testing accuracies with the reduction numbers 16, 8, 4, 2, and 1 are 97.92%, 98.12%, 98.90%, 98.75%, and 98.36%, respectively. This shows that the best number is four, which is more suitable for DAS data. Therefore, we set this hyperparameter to four in subsequent experiments.

3.3. Performance Evaluation of the Proposed ResNet+CBAM

In this section, five models are compared to evaluate the recognition performances, including the 1-D CNN [33], 2-D CNN, ResNet, 2-D CNN+CBAM, and the proposed attention-enhanced ResNet network (ResNet+CBAM). The structure of compared 2-D CNN+CBAM is illustrated in Figure 7. Through the validation set, the optimal parameters of the models (the constructed ResNet+CBAM and the compared 2-D CNN+CBAM) are obtained, as shown in Tables 3 and 4, and those for ResNet and 2-D CNN are obtained by removing CBAMs in the networks of ResNet+CBAM and 2-D CNN+CBAM.

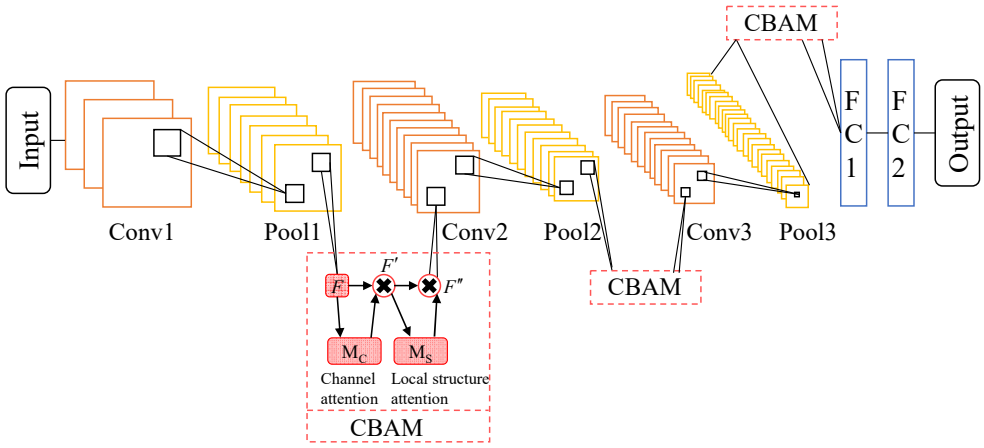


Figure 7. The compared 2-D CNN+CBAM configuration.

Firstly, the average loss curves of cross-entropy for the five models in the training process are comparatively obtained, as detailed in Figure 8a,b. This shows that, in terms of the convergence rate, ResNet is better than 2-D CNN, there is little difference between 2-D CNN and 1-D CNN, the two networks with CBAMs are both better than those without them,

and ResNet+CBAM is the best. Further, the training and validation processes are compared as in Figure 8c,d. From the training accuracy and validation accuracy in Figure 8a,b, ResNet+CBAM has the best training accuracy, which is better than 2-D CNN+CBAM, ResNet, 2-D CNN, and 1-D CNN, successively; and 2-D CNN is the first to be overfitted, followed by 1-D CNN, 2-D CNN+CBAM, and ResNet, while ResNet+CBAM is the last. This highlights the better anti-overfitting ability of the residual blocks in ResNet and ResNet+CBAM.

Table 3. Structural parameters of the ResNet+CBAM network.

Layers	Kernel Size/Stride/Padding	Input Size
Conv1	$1 \times 5 \times 5/1 \times 1$	$50 \times 100 \times 1$
Pool1	$1 \times 25 \times 32/1 \times 25$	$16 \times 46 \times 96$
Residual block1 + CBAM	$1 \times 3 \times 3/1 \times 1/1 \times 1$ $1 \times 3 \times 3/1 \times 1/1 \times 1$	$16 \times 46 \times 96$
Conv2	$1 \times 5 \times 5/1 \times 1$	$16 \times 23 \times 48$
Pool2	$1 \times 25 \times 32/1 \times 25$	$32 \times 19 \times 44$
Residual block2 + CBAM	$1 \times 3 \times 3/1 \times 1/1 \times 1$ $1 \times 3 \times 3/1 \times 1/1 \times 1$	$32 \times 19 \times 44$
FC1	In features = 6336, out features = 144	$32 \times 9 \times 22 = 6336$
FC2	In features = 144, out features = 4	1×144

Table 4. Structural parameters of the 2-D CNN+CBAM network.

Layers	Kernel Size/Stride/Padding	Input Size
Conv1	$3 \times 25 \times 32/1 \times 1/1 \times 12$	$50 \times 100 \times 1$
Pool1	$4 \times 4 \times 32/2 \times 2$	$50 \times 100 \times 32$
CBAM1	$1 \times 3 \times 3/1 \times 1/1 \times 1$ $1 \times 3 \times 3/1 \times 1/1 \times 1$	$50 \times 100 \times 32$
Conv2	$3 \times 25 \times 64/1 \times 1/1 \times 12$	$24 \times 49 \times 32$
Pool2	$4 \times 4 \times 64/2 \times 2$	$24 \times 49 \times 64$
CBAM2	$1 \times 3 \times 3/1 \times 1/1 \times 1$ $1 \times 3 \times 3/1 \times 1/1 \times 1$	$24 \times 49 \times 64$
Conv3	$3 \times 25 \times 96/1 \times 1/1 \times 12$	$4 \times 10 \times 64$
Pool3	$4 \times 4 \times 96/2 \times 2$	$4 \times 10 \times 96$
CBAM3	$1 \times 3 \times 3/1 \times 1/1 \times 1$ $1 \times 3 \times 3/1 \times 1/1 \times 1$	$4 \times 10 \times 96$
FC1	In features = 3840, out features = 200	$4 \times 10 \times 96 = 3840$
FC2	In features = 200, out features = 4	1×200

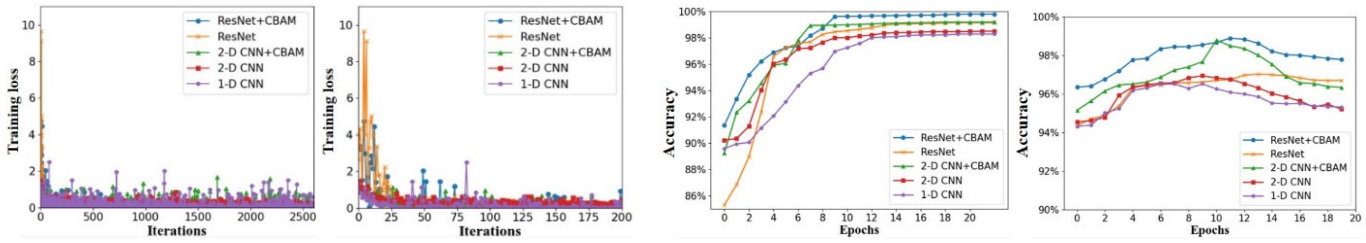


Figure 8. Training loss comparison of the five models: (**a**) the whole curve, (**b**) the enlarged part, and comparison of the five models in the processes of (**c**) training and (**d**) validation.

In order to optimize the structure and hyperparameters of the model and to evaluate the generalization ability of the five models, the 10-fold cross-validation method is used to verify the models. The specific steps of the 10-fold cross-validation experiment are as follows: Firstly, the training and validation sets in Table 2 are added up and then randomly divided into 10 equal subsets of samples. The 10 subsets are traversed successively, the

current subset is taken as the validation set each time, and all the other subsets are taken as a training set to train and evaluate the model. Finally, the average value of 10 evaluation indexes is taken as the final evaluation index. The ten-fold cross-validation of the five models is illustrated in Figure 9. In Figure 9, ResNet+CBAM behaves the best, and the average accuracy is 99.014% for the four events in Table 2, which indicates that the generalization ability of this model is superior to the other four models.

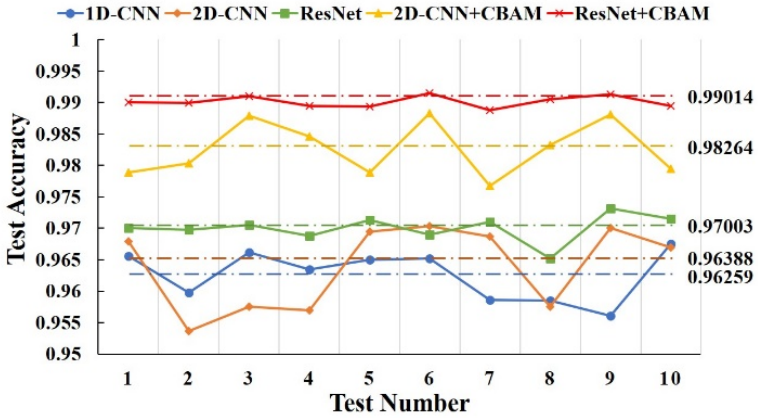


Figure 9. Ten-fold cross-validation of the five models.

After training and cross-validation, we selected the optimal model of each of the five models and their training parameters for the test phase to test the data in the test set. The confusion matrices, ROC curves, and the calculated performance indices are illustrated in Figures 10 and 11, and Table 5. It can be seen that ResNet+CBAM and 2-D CNN+CBAM are both better than ResNet and the 2-D CNN without the CBAMs, which means that attention plays an important role, and it can improve the performance of the network further. With a CBAM, the average recognition accuracy of the 2-D CNN increases from 96.98% to 98.79%, increasing by 1.81%, and the convergence rate is improved according to Figure 8. Similarly, with a CBAM, the average recognition accuracy of ResNet increases from 97.11% to 98.89%, by 1.78%, and the convergence rate also improves. The performance of the two networks with the attention mechanism is significantly better than those without it. Finally, with the same epoch number and reduction number, the average recognition accuracy of ResNet+CBAM is achieved at 98.89%, which is better than the 98.79% of 2-D CNN+CBAM, while the convergence rate is much better than the 2-D CNN+CBAM. It can also be seen from the ROC curve of the test phase that ResNet+CBAM has the largest AUC area of 0.9870, which is better than the other four models, 2-D CNN+CBAM, ResNet, 2-D CNN, and 1-D CNN. In addition, ResNet is slightly better than the 2-D CNN, and then the 1-D CNN, which also indicates that 2-D CNN features extracted from the time-frequency spectrograms are more comprehensive than the 1-D CNN features only extracted in the time domain.

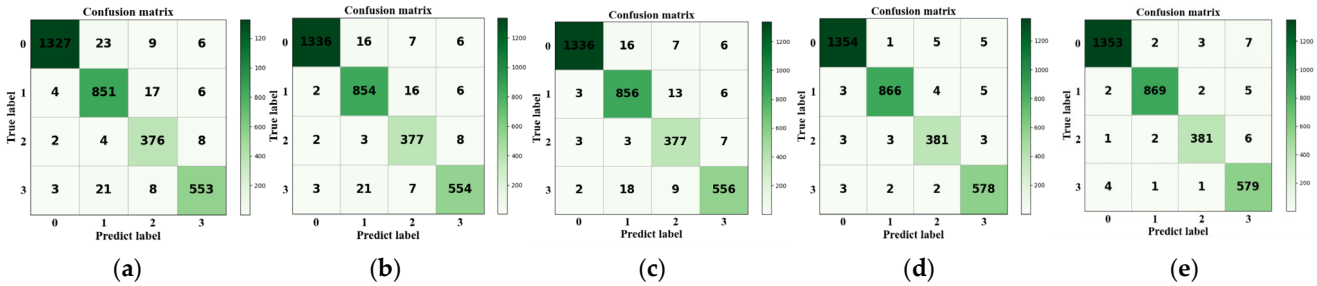


Figure 10. Confusion matrices for (**a**) 1D-CNN, (**b**) 2D-CNN, (**c**) ResNet, (**d**) 2D-CNN+CBAM, and (**e**) ResNet+CBAM.

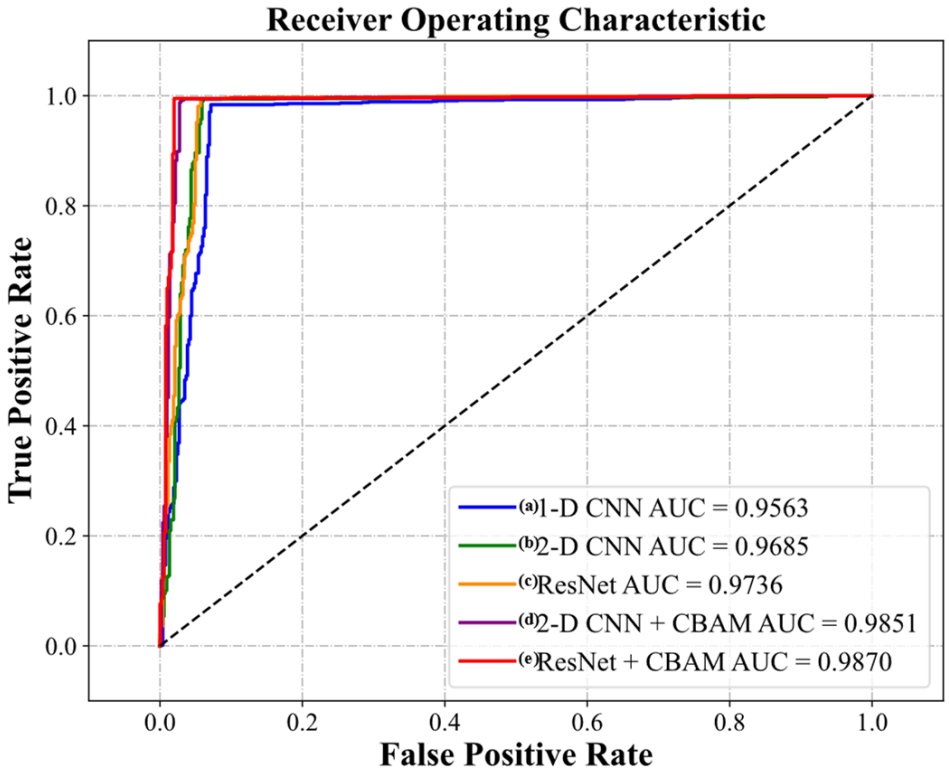


Figure 11. The ROC curves for (**a**) 1D-CNN, (**b**) 2D-CNN, (**c**) ResNet, (**d**) 2D-CNN+CBAM, and (**e**) ResNet+CBAM.

Table 5. Comparison of the recognition results for the five models.

Model	Label	Precision	Recall	F1-Score	Accuracy
1-D CNN [33]	0	0.9875	0.9861	0.9905	0.9655
	1	0.9537	0.9684	0.9775	
	2	0.9553	0.9538	0.9372	
	3	0.9640	0.9351	0.9476	
2-D CNN	0	0.9947	0.9788	0.9867	0.9698
	1	0.9553	0.9727	0.9639	
	2	0.9263	0.9667	0.9461	
	3	0.9652	0.9470	0.9560	
ResNet	0	0.9941	0.9802	0.9871	0.9711
	1	0.9586	0.9749	0.9667	
	2	0.9332	0.9667	0.9497	
	3	0.9670	0.9504	0.9586	
2-D CNN+CBAM	0	0.9934	0.9919	0.9926	0.9879
	1	0.9931	0.9644	0.9785	
	2	0.9719	0.9769	0.9744	
	3	0.9780	0.9880	0.9830	
ResNet+CBAM	0	0.9949	0.9912	0.9930	**0.9889**
	1	0.9943	0.9897	0.9920	
	2	0.9845	0.9769	0.9807	
	3	0.9698	0.9897	0.9796	

In addition, it can be seen from Figure 10 that traffic interference (Label 1) and mechanical excavation (Label 3) are easily confused. The reason may be that the vibration sources of traffic interference and mechanical excavators are both time-varying, and in some time periods, their signals are difficult to distinguish. As such, the recognition of these two types of events is more challenging.

3.4. The Computation Efficiency of the Proposed Method

The time cost of the five models for one test sample on average is compared in Figure 12. It shows the recognition speed of ResNet is faster than 1-D CNN, 2-D CNN, ResNet+CBAM, and 2-D CNN+CBAM, successively. In more detail, the 1-D CNN is faster than the 2-D CNN because of the simpler structure, and the recognition time of ResNet is 1.5 ms, which is faster than the 2.5 ms of the 2-D CNN for one sample. Furthermore, due to the addition of the attention mechanism module (CBAM), the recognition times of 2-D CNN+CBAM and ResNet+CBAM are 18.5 ms and 3.3 ms, which are longer than those of 2-D CNN and ResNet. At any rate, the proposed ResNet+CBAM has obvious advantages in online real-time processing, which are important in practical applications.

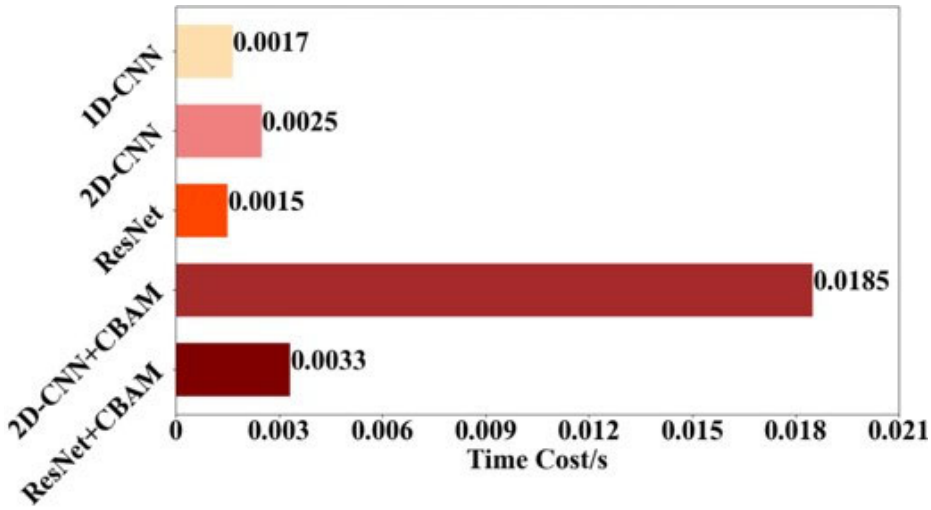


Figure 12. Time cost comparison of the five models for one sample on average. Notes: The test is on a commercial processor (GPU: GTX1080Ti, CPU: Intel i7 dual-core, Memory: 32G).

3.5. The Challenging Test Case in Fields

When dealing with different scenes and inconsistent laying or burying environments, a network's generalization and robustness are important. Thus, two challenging data sets are further constructed, in which only the inconsistent and atypical samples in the database are selected and tested. Since it has the worst recognition performance and its one-dimensional time series contains less feature information, 1-D CNN is excluded from the test. The other models are compared with the data demonstrated in Table 6.

Table 6. Comparison of the recognition results for the five models.

Event Type	Typical and Inconsistent	Atypical and Inconsistent
Label 0	250	250
Label 1	250	250
Label 2	250	250
Label 3	250	250

Here the "typical and inconsistent" data set is composed of balanced—but typical and inconsistent—samples selected by hand, which denote the inconsistent samples of the same type of test events. The inconsistent signals mainly refer to the signals whose amplitudes are inconsistent due to the different acting forces or the varying buried condition, but the signal-changing law in the time domain is basically the same or not distorted. Furthermore, the "atypical and inconsistent" data set contains only atypical and inconsistent samples, which mainly refer to signals that have a distorted shape or that have an inconsistent evolution law from the perspective of the human eyes due to the different acting period or

other unpredictable interfering factors for the same type of event. The inconsistent data sets stand for extremely challenging cases.

The test results for the two inconsistent data sets are compared in Figure 13. It can be seen that ResNet's average recognition performance is superior to that of 2-D CNN in both the typical and inconsistent data sets and the atypical and inconsistent data set, while the two networks with the attention modules are better than those without them, and the four performance indices can be achieved at more than 91.08% for ResNet+CBAM, which is better than ResNet, and then the 2-D CNN+CBAM and then the 2-D CNN. On the whole, the results of the atypical and inconsistent data set are worse than those of the typical and inconsistent data set for all the models, which is consistent with the field complexity in the two actual cases. It also shows that the proposed ResNet+CBAM has the best generalization capability and generality in this challenging case.

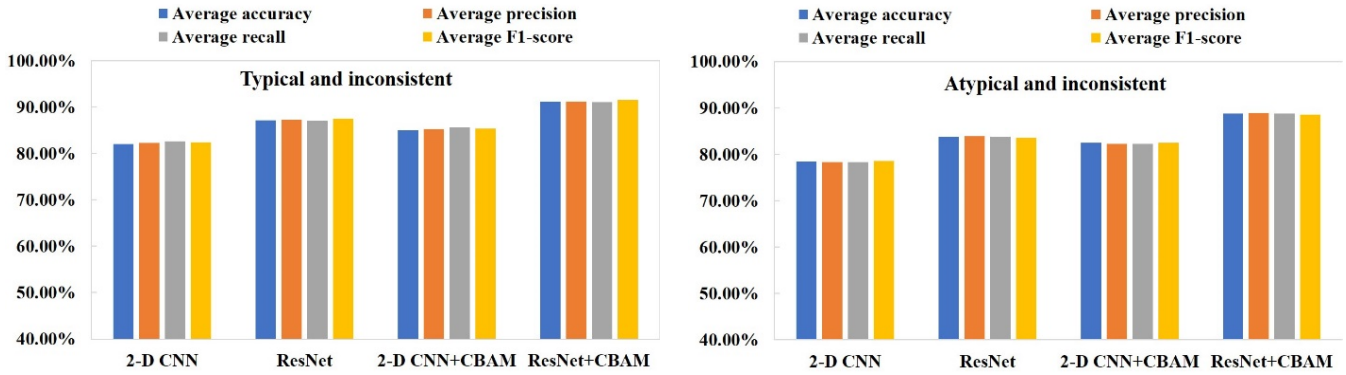


Figure 13. Performance comparison of the 2-D CNN, ResNet, 2-D CNN+CBAM, and ResNet+CBAM with the (**a**) typical and inconsistent and (**b**) atypical and inconsistent data sets in Table 6.

4. Conclusions

In this paper, a novel recognition method is proposed for DAS by using an end-to-end, attention-enhanced ResNet model. The effectiveness of different time-frequency features of STFT and MFCC are compared and the better one is chosen as the input of the network. The field test results show that the proposed ResNet+CBAM model behaves the best in recognition accuracy, convergence rate, generalization capability, and computational efficiency among the five models, namely 1-D CNN, 2-D CNN, ResNet, 2-D CNN+CBAM, and ResNet+CBAM. In particular, its generalization ability and time efficiency are quite exciting, which could be very promising in online, long-distance, distributed monitoring applications. In the future, small training samples or imbalanced data sets will be tested, which is also challenging in practice.

Author Contributions: Conceptualization, X.L. and H.W.; data curation, X.L.; formal analysis, X.L., Y.W. (Yufeng Wang) and G.Y.; funding acquisition, H.W., Y.S. (Yuanfeng Song) and Y.W. (Yu Wu); investigation, X.L., Y.T., Y.S. (Yuwen Sun) and L.L.; methodology, X.L.; project administration, H.W.; resources, X.L., H.W., Y.W. (Yufeng Wang), Y.T., Y.S. (Yuwen Sun), L.L. and Y.W. (Yu Wu); software, X.L.; supervision, Y.S. (Yuanfeng Song), Y.W. (Yu Wu) and G.Y.; validation, X.L.; visualization, X.L.; writing—original draft, X.L.; writing—review & editing, X.L. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of China (Grant No. U21A20453, Grant No. 41527805, Grant No. 61290312, and Grant No. 61301275), and supported by the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT, IRT1218) and the 111 Project (B14039).

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

References

1. Taylor, H.F.; Juarez, J.C. Distributed fiber-optic intrusion sensor system. *J. Lightwave Technol.* **2005**, *23*, 2081–2087.
2. Wang, Z.; Zeng, J.; Li, J.; Peng, F.; Zhang, L.; Zhou, Y.; Wu, H.; Rao, Y. 175km phase-sensitive OTDR with hybrid distributed amplification. In Proceedings of the 23rd International Conference on Optical Fibre Sensors, Santander, Spain, 2–6 June 2014; Volume 9157, pp. 1562–1565.
3. Wu, H.; Qian, Y.; Zhang, W.; Tang, C. Feature extraction and identification in distributed optical-fiber vibration sensing system for oil pipeline safety monitoring. *Photonic Sens.* **2017**, *7*, 305–310. [CrossRef]
4. Lyu, A.; Li, J. On-line monitoring system of 35 kV 3-core submarine power cable based on φ-OTDR. *Sens. Actuators A Phys.* **2018**, *273*, 134–139.
5. Wu, H.; Wang, Z.; Peng, F.; Peng, Z.; Li, X.; Yu, W.; Rao, Y. Field test of a fully distributed fiber optic intrusion detection system for long-distance security monitoring of national borderline. In Proceedings of the 23rd International Conference on Optical Fiber Sensors, Santander, Spain, 2–6 June 2014; Volume 9157, pp. 1285–1288.
6. Williams, E.; Fernández-Ruiz, M.; Magalhaes, R.; Vanthillo, R.; Zhan, Z.; González-Herráez, M.; Martins, H.F. Distributed sensing of microseisms and teleseisms with submarine dark fibers. *Nat. Commun.* **2019**, *10*, 5778. [CrossRef] [PubMed]
7. Rao, Y.; Wang, Z.; Wu, H.; Ran, Z.; Han, B. Recent advances in phase-sensitive optical time domain reflectometry (Φ-OTDR). *Photonic Sens.* **2021**, *11*, 1–30. [CrossRef]
8. Dong, Y.; Chen, X.; Liu, E.; Fu, C.; Zhang, H.; Lu, Z. Quantitative measurement of dynamic nanostrain based on a phase-sensitive optical time domain reflectometer. *Appl. Opt.* **2016**, *55*, 7810. [CrossRef] [PubMed]
9. Wang, B.; Ba, D.; Chu, Q.; Qiu, L.; Zhou, D.; Dong, Y. High-sensitivity distributed dynamic strain sensing by combining Rayleigh and Brillouin scattering. *Opto-Electron. Adv.* **2020**, *3*, 200013. [CrossRef]
10. Zheng, H.; Zhang, J.; Guo, N.; Zhu, T. Distributed optical fiber sensor for dynamic measurement. *J. Lightwave Technol.* **2021**, *39*, 3801–3811. [CrossRef]
11. Shao, L.Y.; Liu, S.; Bandyopadhyay, S.; Yu, F.; Xu, W.; Wang, C.; Zhang, J. Data-driven distributed optical vibration sensors: A review. *IEEE Sens. J.* **2019**, *20*, 6224–6239. [CrossRef]
12. Lu, Y.; Tao, Z.; Liang, C.; Bao, X. Distributed vibration sensor based on coherent detection of Phase-OTDR. *J. Lightwave Technol.* **2010**, *28*, 3243–3249.
13. Zhu, T.; Xiao, X.; He, Q.; Diao, D. Enhancement of SNR and spatial resolution in phi-OTDR system by using two-dimensional edge detection method. *J. Lightwave Technol.* **2013**, *31*, 2851–2856. [CrossRef]
14. Qin, Z.; Chen, L.; Bao, X. Wavelet denoising method for improving detection performance of distributed vibration sensor. *IEEE Photonics Technol. Lett.* **2012**, *24*, 542. [CrossRef]
15. Hui, X.; Zheng, S.; Zhou, J.; Hao, C.; Jin, X.; Zhang, X. Hilbert-Huang transform time-frequency analysis in phi-OTDR distributed sensor. *IEEE Photonics Technol. Lett.* **2014**, *26*, 2403–2406. [CrossRef]
16. Wu, H.; Wang, J.; Wu, X.; Wu, Y. Real intrusion detection for distributed fiber fence in practical strong fluctuated noisy backgrounds. *Sens. Lett.* **2012**, *10*, 1557–1561. [CrossRef]
17. Wu, H.; Xiao, S.; Li, X.; Wang, Z.; Xu, J.; Rao, Y. Separation and determination of the disturbing signals in Phase-sensitive optical time domain reflectometry (Φ-OTDR). *J. Lightwave Technol.* **2015**, *33*, 3156–3162. [CrossRef]
18. Liu, S.; Yu, F.; Hong, R.; Xu, W.; Shao, L.; Wang, F. Advances in phase-sensitive optical time-domain reflectometry. *Opto-Electron. Adv.* **2022**, *5*, 200078. [CrossRef]
19. Tan, D.; Tian, X.; Wei, S.; Yan, Z.; Hong, Z. An oil and gas pipeline pre-warning system based on Φ-OTDR. In Proceedings of the 23rd International Conference on Optical Fibre Sensors, Santander, Spain, 2–6 June 2014; Volume 9157, pp. 1269–1272.
20. Zhu, H.; Pan, C.; Sun, X. Vibration waveform reproduction and location of OTDR based distributed optical-fiber vibration sensing system. In *Quantum Sensing and Nanophotonic Devices XI, Proceedings of the International Society for Optics and Photonics, San Francisco, CA, USA, 1–6 February 2014*; SPIE: Bellingham, WA, USA, 2014; Volume 8993, pp. 277–282.
21. Fang, N.; Wang, L.; Jia, D.; Shan, C.; Huang, Z. Walking intrusion signal recognition method for fiber fence system. In Proceedings of the 2009 Asia Communications and Photonics conference and Exhibition (ACP), Shanghai, China, 2–6 November 2009; Volume 52, pp. 2381–2384.
22. Wang, Z.; Pan, Z.; Ye, Q.; Cai, H.; Qu, R.; Fang, Z. Fast pattern recognition based on frequency spectrum analysis used for intrusion alarming in optical fiber fence. *Chin. J. Lasers* **2015**, *42*, 0405010. [CrossRef]
23. Sun, Q.; Feng, H.; Yan, X.; Zeng, Z. Recognition of a phase-sensitivity OTDR sensing system based on morphologic feature extraction. *Sensors* **2015**, *15*, 15179–15197. [CrossRef]
24. Tu, G.; Yu, B.; Zhen, S.; Qian, K.; Zhang, X. Enhancement of signal identification and extraction in a Φ-OTDR vibration Sensor. *IEEE Photonics J.* **2017**, *9*, 1–10. [CrossRef]
25. Tejedor, J.; Martins, H.F.; Piote, D.; Macias-Guarasa, J.; Pastor-Graells, J.; Martin-Lopez, S.; González-Herráez, M. Toward prevention of pipeline integrity threats using a smart fiber-optic surveillance system. *J. Lightwave Technol.* **2016**, *34*, 4445–4453. [CrossRef]
26. Wang, B.; Pi, S.; Sun, Q.; Jia, B. Improved wavelet packet classification algorithm for vibrational intrusions in distributed fiber-optic monitoring systems. *Opt. Eng.* **2015**, *54*, 055104. [CrossRef]
27. Tian, Q.; Zhao, C.; Zhang, Y.; Qu, H. Intrusion signal recognition in OFPS under multi-level wavelet decomposition based on RVFL neural network. *Optik* **2017**, *146*, 38–50. [CrossRef]

28. Fedorov, A.K.; Anufriev, M.N.; Zhirnov, A.A.; Stepanov, K.V.; Nesterov, E.T.; Namiot, D.E.; Pnev, A.B. Note: Gaussian mixture model for event recognition in optical time-domain reflectometry based sensing systems. *Rev. Sci. Instrum.* **2016**, *87*, 036107. [CrossRef] [PubMed]
29. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A new convolutional neural network based data-driven fault diagnosis method. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5990–5998. [CrossRef]
30. Xu, C.; Guan, J.; Bao, M.; Lu, J.; Ye, W. Pattern recognition based on time-frequency analysis and convolutional neural networks for vibrational events in φ-OTDR. *Opt. Eng.* **2018**, *57*, 016103. [CrossRef]
31. Aktas, M.; Akgun, T.; Demircin, M.U.; Buyukaydin, D. Deep learning based multi-threat classification for phase-OTDR fiber optic distributed acoustic sensing applications. *Fiber Opt. Sens. Appl.* **2017**, *10208*, 102080.
32. Jiang, F.; Li, H.; Zhang, Z.; Zhang, X. An event recognition method for fiber distributed acoustic sensing systems based on the combination of MFCC and CNN. In Proceedings of the International Conference on Optical Instruments and Technology: Advanced Optical Sensors and Applications, Beijing, China, 28–30 October 2017; Volume 10618, pp. 15–21.
33. Wu, H.; Chen, J.; Liu, X.; Xiao, Y.; Wang, M.; Zheng, Y.; Rao, Y. One-dimensional CNN-based intelligent recognition of vibrations in pipeline monitoring with DAS. *J. Lightwave Technol.* **2019**, *37*, 4359–4366. [CrossRef]
34. Wu, H.; Wang, C.; Liu, X.; Gan, D.; Liu, Y.; Rao, Y.; Olaribigbe, A.O. Intelligent target recognition for distributed acoustic sensors by using both manual and deep features. *Appl. Opt.* **2021**, *60*, 6878–6887. [CrossRef]
35. Shi, Y.; Wang, Y.; Zhao, L.; Fan, Z. An event recognition method for Φ-OTDR sensing system based on deep learning. *Sensors* **2019**, *19*, 3421. [CrossRef]
36. Sun, Z.; Liu, K.; Jiang, J.; Xu, T.; Wang, S.; Guo, H.; Liu, T. Optical Fiber Distributed Vibration Sensing Using Grayscale Image and Multi-Class Deep Learning Framework for Multi-Event Recognition. *IEEE Sens. J.* **2021**, *21*, 19112–19120. [CrossRef]
37. Yang, Y.; Li, Y.; Zhang, H. Pipeline Safety Early Warning Method for Distributed Signal using Bilinear CNN and LightGBM. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; 2021; pp. 4110–4114.
38. Wu, J.; Guan, L.Y.; Bao, M.; Xu, Y.; Ye, W. Vibration events recognition of optical fiber based on multi-scale 1-D CNN. *Opto-Electron. Eng.* **2019**, *46*, 180493.
39. Wu, H.; Yang, S.; Liu, X.; Xu, C.; Lu, H.; Wang, C.; Olaribigbe, A.O. Simultaneous extraction of multi-scale structural features and the sequential information with an end-to-end mCNN-HMM combined model for DAS. *J. Lightwave Technol.* **2021**, *39*, 6606–6616. [CrossRef]
40. Chen, X.; Xu, C. Disturbance pattern recognition based on an ALSTM in a long-distance φ-OTDR sensing system. *Microw. Opt. Technol. Lett.* **2019**, *62*, 1002. [CrossRef]
41. Sun, M.; Yu, M.; Lv, P.; Li, A.; Wang, H.; Zhang, X.; Zhang, T. Man-made Threat Event Recognition Based on Distributed Optical Fiber Vibration Sensing and SE-WaveNet. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1. [CrossRef]
42. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-Attentional Convolutional Neural Network for Scene Text Detection. *IEEE Trans. Image Process.* **2016**, *25*, 2529–2541. [CrossRef]
45. Ferreira, J.F.; Dias, J. Attentional Mechanisms for Socially Interactive Robots–A Survey. *IEEE Trans. Auton. Ment. Dev.* **2014**, *6*, 110–125. [CrossRef]
46. Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; Maybank, S. Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2016; pp. 4854–4863.
47. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
48. Masoudi, A.; Belal, M.; Newson, T.P. A distributed optical fiber dynamic strain sensor based on phase-OTDR. *Meas. Sci. Technol.* **2013**, *24*, 085204.
49. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
50. Zhang, C.; Zhu, L.; Yu, L. The attention mechanism in the convolutional neural network review. *Comput. Eng. Appl.* **2021**, *57*, 9.