

EMPIRICAL RESEARCH

Open Access



Attention mechanism combined with residual recurrent neural network for sound event detection and localization

Chaofeng Lan^{1†}, Lei Zhang², Yuanyuan Zhang^{1†}, Lirong Fu^{3*} , Chao Sun¹, Yulan Han¹ and Meng Zhang^{4†}

Abstract

In the task of sound event detection and localization (SEDL) in a complex environment, the acoustic signals of different events usually have nonlinear superposition, so the detection and localization effect is not good. Given this, this paper is based on the Residual-spatially and channel Squeeze-Excitation (Res-scSE) model. Combined with Multiple-scale Convolutional Recurrent Neural Network (M-CRNN), the Res-scSE-CRNN model is proposed. Firstly, to solve the problem of insufficient extraction of time-frequency feature in single-size convolution kernel, multi-scale feature fusion is carried out by using the feature hierarchy of the convolutional neural network to improve the accuracy of detection. Secondly, aiming at the problem of overlapping audio event localization accuracy is not high, with Res-scSE to replace common convolution module and add residual structure to strengthen the feature extraction, and combining with an attention mechanism to enhance neural network channels and spatial relationships, to improve the network to extract the characteristics of directivity, achieve the goal of the overlapped audio localization. In this paper, experiments are carried out in the open dataset DCASE2019, and evaluation indicators are used to analyze the effectiveness of the proposed model and baseline model in the detection and localization of audio events. The results show that compared with the M-CRNN model, the detection error rate of Res-scSE-CRNN model is reduced 4%, the F1-Score is increased 3.4%, the localization error is reduced by 22.8°, and the frame recall rate is increased 3%.

Keywords: Sound event, Detection and localization, Convolutional cyclic neural network, Multi-scale feature fusion, Space channel squeeze excitation module

1 The introduction

Sound as an important medium of information transmission has always been the focus of the majority of researchers. Sound is penetrating; it is not affected by light and can navigate around obstacles as it travels. The system based on SEDL has the advantages of small size, and high security in intelligent medical care, smart home,

and security monitoring [1] aspect has a wide application prospect.

SEDL contains two tasks. One is to classify and identify sound events in a specific environment and mark the start and end time of each sound event to realize sound event detection (SED). The other is to estimate the corresponding spatial localization trajectory of sound events in a specific environment, namely directions of arrival (DOA) azimuth and elevation. In the early days of SED research in the 1920s and 1980s, SED algorithms based on traditional machine learning dominated, Mesaros et al. [2] modeled various audio events through Hidden Markov Model(HMM) and then locate the start and end times of audio events. Although this method is effective, the decoding calculation is huge. Therefore, with

[†]Chaofeng Lan, Yuanyuan Zhang and Meng Zhang contributed equally to this work.

*Correspondence: 993560@hainanu.edu.cn

³ Mechanical and Electrical Engineering College, Hainan University, 570228 Haikou, People's Republic of China
Full list of author information is available at the end of the article

Nonnegative Matrix Factorization(NMF) [3] and random forest [4] the SED method can reduce the amount of computation and achieve good results in the early small sample set. However, in traditional machine learning methods, complex feature engineering is often needed to extract audio signal features.

In the 21st century, the emergence of deep learning technology brings new development opportunities for SEDL. Literature [5] proposes that the Convolutional Neural Network (CNN) learning framework can automatically learn features from audio sample data, which overcomes the disadvantage of manual extraction of audio features in traditional machine learning. VU et al. [6] proposed that a Recurrent Neural Network(RNN) is used to learn the temporal correlation of environmental sound signals, and to model the short-term and long-term dependence of time frames, which has achieved good results. Song Jiannan et al. [7] built a binarization Network system based on Deep Neural Network (DNN) for SED, which reduces the occupation of computer resources and has a good detection effect. This is because the CNN-based method cannot effectively capture the long time dependence of audio segments in SED tasks. To solve this problem, Cakir et al. [8] proposed a Convolutional Recurrent Neural Network (CRNN) obtained by combining CNN and RNN for multi-sound event detection, and the powerful feature extraction ability of CNN can be used to extract frequency invariant features. Bidirectional RNN is used to obtain the long-time information about sound events, and the detection performance is better. Literature [9] proposed a three-dimensional CRNN model to extract Generalized Cross Correlation (GCC) variation features between multi-channel audio pairs and learn interchannel features, to identify overlapping audio events more quickly and achieve better SED effects in less training time. Compared with mono audio, the performance of the model is significantly improved. Kong Q et al. [10] also proposed a CRNN model based on a time-frequency segmentation framework to train weak label data to reduce the influence of weak label datasets on poor model detection. Therefore, the CRNN model became the most widely used model in the SED field at that time.

As neural networks based on attention mechanisms are widely used in text classification and emotion classification tasks, some researchers also introduce attention mechanisms into SED tasks, such as Turab et al. [11] combined the latest capsule Network (CapsNet) model and the attention mechanism to learn the most significant features of audio signals to detect large-scale weak label audio events, and a breakthrough was achieved in DCASE. Liping Yang et al. [12] proposed a sound event detection method (ATCC-CRNN) to solve the CRNN

sound event detection model trained in an end-to-end manner cannot constrain the functions of CNN and RNN structures functionally. The results show that the proposed ATCC-CRNN method promotes the function division of the CRNN model and improves the generalization ability of the CRNN sound event detection model effectively.

Yuzhan Huang et al. [13] proved that neither simple wavelet transform nor simple Kalman filter can satisfy the demand of denoising in the case of uncertain noise distribution. Many researchers have introduced multi-scale methods based on attention and achieved good research results in SED, such as Lee J, et al. [14] proposed using the combination of multi-level and multi-scale features is highly effective in music auto-tagging because of the music auto-tagging is distinguished from image classification in that the tags are highly diverse and have different levels of abstraction. The method outperforms the previous state-of-the-art methods on the MagnaTagATune dataset and the Million Song Dataset. Chen Xinxing et al. [15] proposed Multi-scale feature fusion and channel weighting methods are proposed to improve the classification performance of SED models effectively. QiuPeng et al. [16] proposed multi-scale attention fusion-based SED method to reduce the influence of noise in the feature of sound time-frequency graph and fuse the feature with multi-size convolution kernel. This network model has higher recognition accuracy than the traditional SED model. Han Xinyuan et al. [17] built Ghost convolution time-frequency segmentation attention network model. Among them, compared with ordinary convolution, Ghost convolution uses fewer parameters and saves computing resources.

Trowitzsch I et al. [18] proposed an approach that robustly binds localization with the detection of sound events in a binaural robotic system. The use of spatial stream segregation which produces probabilistic time-frequency masks for individual sources attributable to separate localizations, enabling segregated sound event detection operating on these streams.

With the wide application of deep learning technology in the field of signal processing, DOA estimation by deep learning technology can be well combined with other tasks. For example, Cao et al. [19] proposed a two-stage strategy which training single SED and DOA models, and DOA outputs are obtained by using SED outputs as masks. This scheme is significantly better than the CRNN baseline model trained together. Ranjan et al. [20] proposed the method of ResNet combined with RNN for SEDL, which improves the detection and localization performance based on avoiding network degradation. Tan et al. [21] proposed the ResNet fusion RNN and delay estimation algorithm for SEDL on this

basis, which not only ensures the detection accuracy, but also reduces the error of localization estimation. In 2020, Naranjo-Alcazar et al. [22] proposed squeeze excitation fusion CNN method for SEDL, which achieves excellent detection and localization results in overlapping audio events.

To sum up, the existing SED methods based on neural networks have higher detection accuracy compared with traditional SED methods, but they involve a large number of parameter calculations, resulting in large storage space and a long forward reasoning time. To reduce the number of parameters of the neural networks and improve the SED performance, a multi-scale convolution fusion RNN method is proposed. The method adopts multi-scale convolution fusion to obtain more advanced features from the time domain and frequency domain of audio, and has higher detection accuracy with less number of parameters. With the deep processing of the input features, the neural network can get more advanced feature outputs with a stronger ability to express information. But at the same time, because the gradient of backward transmission becomes unstable, it will also be accompanied by gradient explosion or gradient dispersion. The performance of the network is reduced and the convergence speed is slow, resulting in a decrease of SEDL accuracy. Because of the above problems, this paper proposes the method of ResNet integrating attention mechanism and RNN to introduce attention mechanism to avoid the loss of key channel features and improve the overall performance of the model under the condition of reducing network degradation.

2 M-CRNN model construction

2.1 Network structure of M-CRNN

CNN performs well in the application of image classification and segmentation. In the case of audio features, it is also a two-dimensional image that is fed into the computer for processing. Therefore, different feature images of sound events are processed by the image processing method, and the processed image features are sent to CNN for training to output the category labels of the events. In addition to CNN, RNN has strong performance in processing sequence data. Gated Recurrent Unit (GRU) is a variant network in RNN that can effectively capture the contextual characteristics of time series and perform better in longer series. The structure of the Bidirectional Gating Recurrent Unit (Bi-GRU) makes the whole GRU more powerful, making the current moment hidden state not only related to the previous moment hidden state but also related to the future moment's hidden state. Extract global context information and output long-term and short-term dependency characteristics. In the SEDL model, the input characteristics are time-series-dependent spectral characteristics. Using the network based on time interdependence, the characteristic information of each moment and its sequence can be trained. Compared with the simple CNN network, the network combined with CNN and RNN has a stronger time information processing ability. The schematic diagram of the CRNN network structure is shown in Fig. 1 [23].

The CRNN network structure in Fig. 1 includes four major parts: The first part is feature extraction, which mainly obtains two groups of different classification

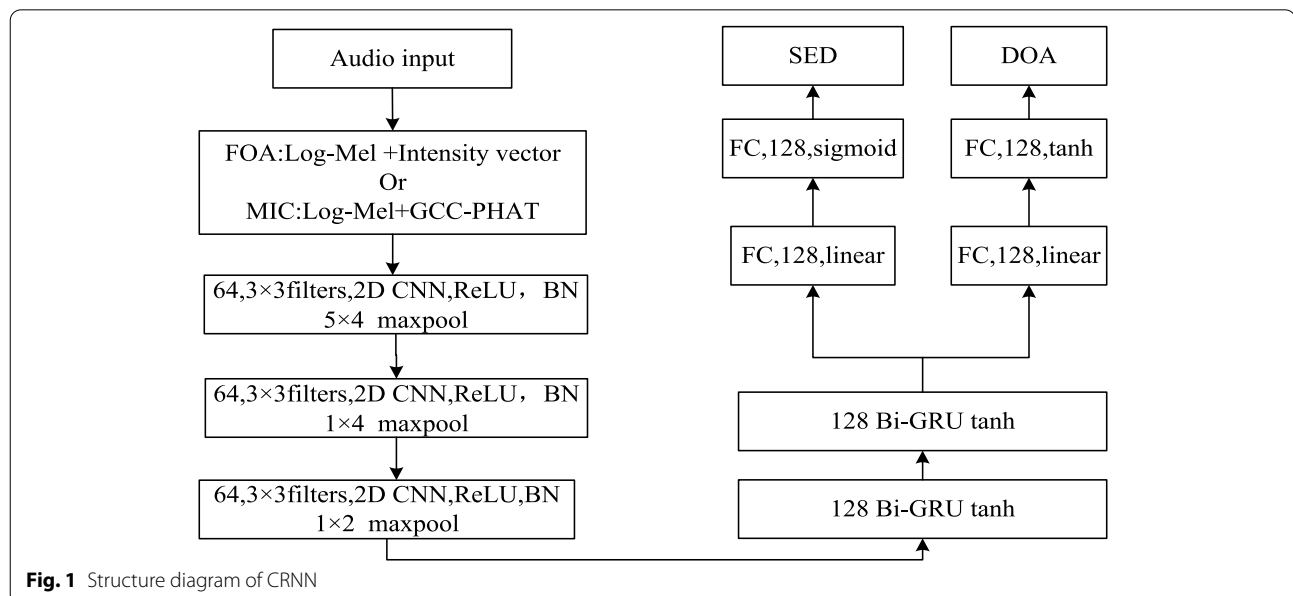


Fig. 1 Structure diagram of CRNN

features from two different datasets, namely, log-mel spectrum and sound intensity vector spectrum from FOA dataset and log-mel spectrum and generalized cross-correlation spectrum from MIC dataset. In the experiment part, only one group can be selected as the feature input of the network layer. The second part is a 3-layer two-dimensional convolution layer, and the convolution kernel of each convolution layer is 3×3 . The convolutional layer learns the displacement invariant features from the input features, and after the non-linear operation of the ReLU activation function, the features are processed by Batch Normalization and max-pooling. The Max Pool of each layer is respectively. The third part is the two-layer GRU layer, which learns the time-frequency structure from the feature whose output is 128 in the upper layer, and obtains deep information input to the next layer network through the nonlinear operation of the Tanh activation function. The fourth part is to learn SED and DOA estimation through two independent branches of two fully connected layers. The SED output layer has a linear activation function and a sigmoid activation function, and the output corresponds to the time activity of C kinds of sound events under time-frequency. Similarly, the DOA output layer also has a linear activation function and Tanh activation function, which output DOA trajectories corresponding to C kinds of acoustic events at the same time resolution.

Since convolution kernels at different scales can learn deep features at different scales, and because of the insufficient feature extraction of single-scale convolution in CRNN, this paper replaced the convolution layer of CRNN with a multi-scale fusion convolution module and

proposed the M-CRNN model, the structure of which is shown in Fig. 2.

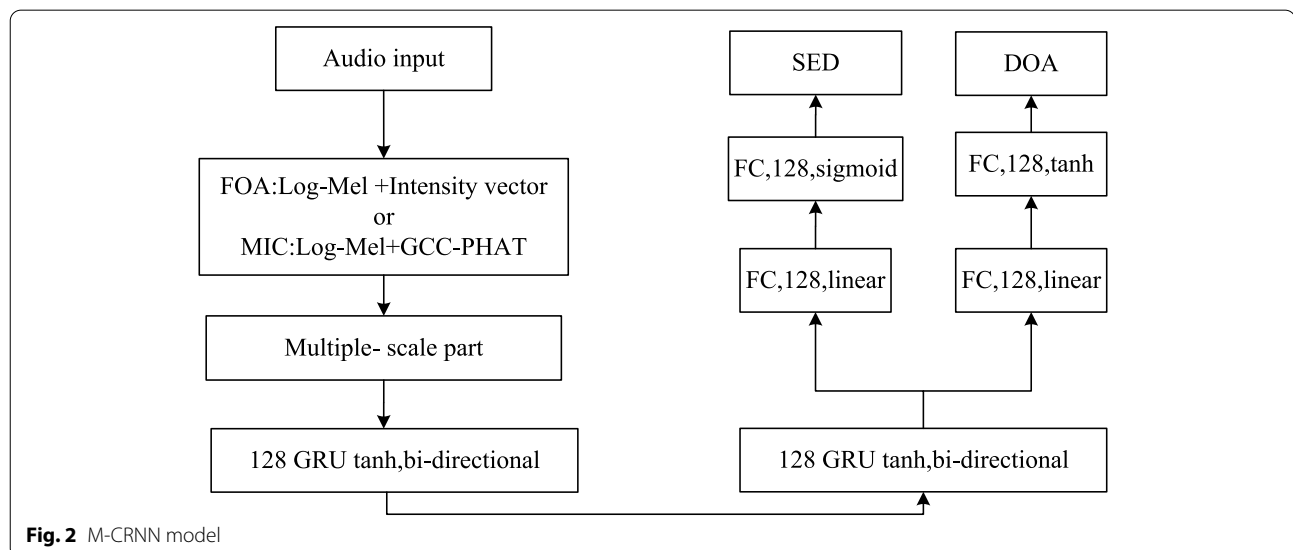
In Fig. 2, four groups of parallel convolution kernels of different scales are used to extract features from the time domain and frequency domain respectively, and the fused features are input into the GRU network. SED and DOA are output after the sigmoid activation function and the tanh activation function.

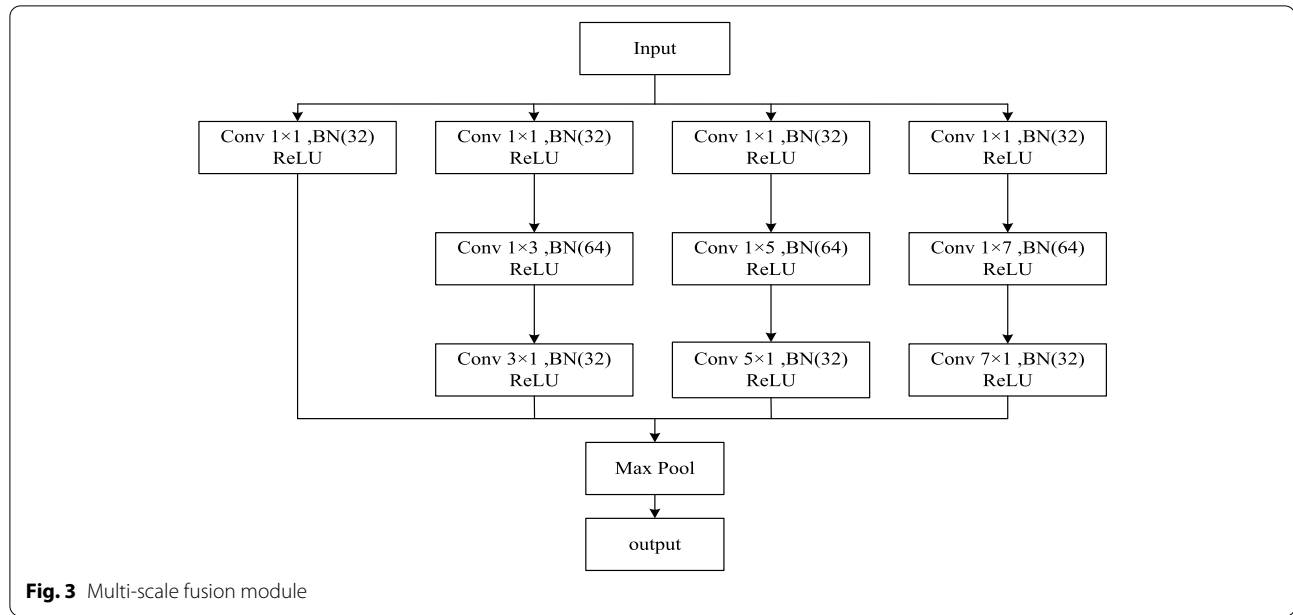
2.2 Multi-scale feature extraction module

Due to the single-scale design limits the capability of model feature extraction to a certain extent, even if the same type of event may have different durations, and the components of sound time and frequency are distributed differently. In order to meet the requirements of complementation and enhancement of features of the same acoustic event at different scales, the structure and network parameter Settings of the multi-scale convolution model proposed in Fig. 2 are shown in Fig. 3.

In Fig. 3, the multi-scale fusion module adopts four sets of parallel convolution branches to extract features from time-frequency graphs. Convolution at different scales is complementary to each other to improve the learning ability of the model for different features. Convolution $n \times n$ is replaced by $1 \times n$, convolution and $n \times 1$ convolution, and n represents the number of convolution kernels, saving the cost of calculation time.

The first group of parallel CNN branches has only one convolution layer, and the convolution kernel size is 1×1 . After convolution, the features are sent to the BN layer for batch training. BN parameter is set to 32, and then output to Max Pool for dimensionality reduction through the nonlinear operation of ReLU activation





function. The second group of parallel CNN branches has three convolution layers, and the convolution kernel size is 1×1 , 1×3 , and 3×1 , respectively. BN is carried out after each convolution operation, and parameter settings are 32, 64, and 32, respectively. The ReLU function is used for activation function. The third group of parallel CNN branches also has three convolution layers, and the convolution kernel size is 1×1 , 1×5 , and 5×1 , respectively. The features are still sent into the BN layer after each convolution operation, and the parameter setting and activation function are the same as those of the second group. The fourth group of parallel CNN branches also has three convolution layers, and the convolution kernel size is 1×1 , 1×7 , and 7×1 , respectively. BN is carried out after each convolution operation, and the parameter setting and activation function are the same as the previous group. Frequency domain features are obtained by frequency domain convolution, and time-domain features by time-domain convolution. 3×1 , $1 \times n$ ($n = 3, 5, 7$) is used to obtain frequency domain features, and n ($n = 3, 5, 7$) $\times 1$ is used to obtain time-domain features.

3 Res-scSE-CRNN model construction

To enhance the expression of channel feature and spatial feature information and avoid network degradation, the attention mechanism fusion module and residual structure are presented in this part. Based on the M-CRNN model in Part 2, the localization performance needs to be improved. The Residual-spatially and channel Squeeze Excitation Recurrent Neural Network(Res-scSE-CRNN)

model is constructed to improve localization performance.

3.1 Residual structure

With the deepening of the network layer, the gradient used for reverse transmission in the network will appear very small with continuous multiplication, resulting in the disappearance of the gradient, and the shallow parameters cannot be updated. In the process of neural network training, it is usually hoped to achieve a better learning effect by choosing deeper network structure than a shallow one. However, with the increase in depth, the model is difficult to learn more than five mapping parameters correctly, and the redundant network layer is easy to cause network degradation [24].

To solve the above problems, this paper proposes to use the characteristics of ResNet, that is, to train the deeper network with fewer parameters, to reduce the problem of gradient disappearance and degradation. The structure of ResNet is shown in Fig. 4.

In Fig. 4, x is the input of ResNet and $F(x)$ the output after linear operation of two convolution layers and non-linear activation of the activation function. The calculation method is as follows:

$$F(x) = w_i \sigma(w_{i-1}x) \quad (1)$$

where w_i is the value of i layer weight, where $i = 2$, $\sigma(x)$ represents ReLU activation function.

It can be seen from Eq. (1) and Fig. 4 that the learning objective becomes $F(x) = H(x) - x$ and it can be seen that when ResNet is used to design the network layer, the network layer is optimized, and the fast connection does

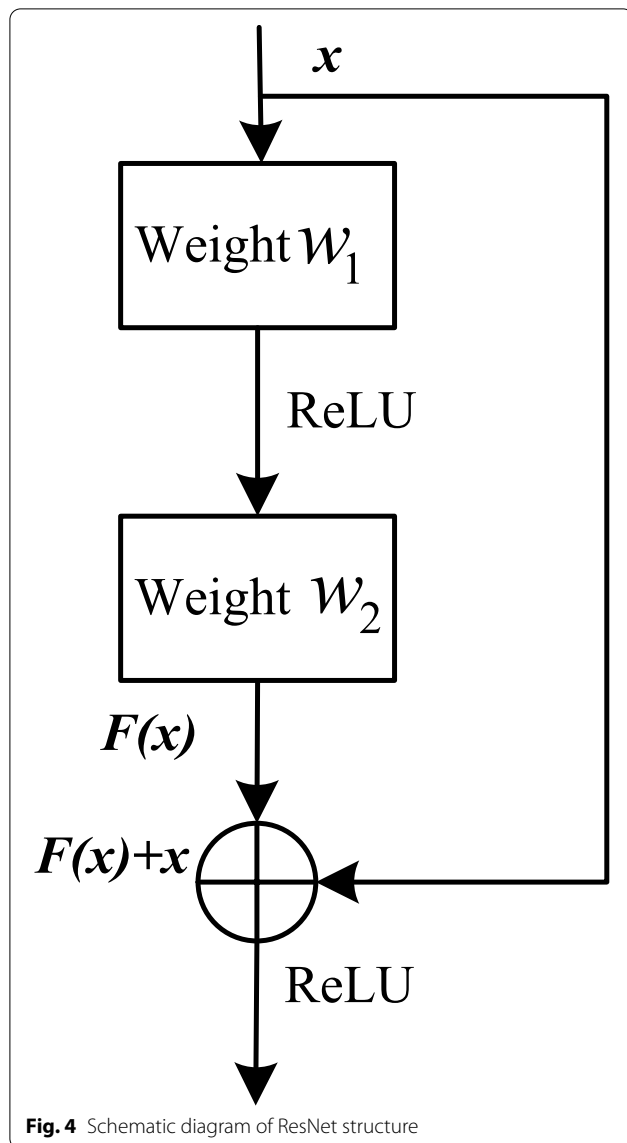


Fig. 4 Schematic diagram of ResNet structure

not add more parameters, nor does it add extra computing cost.

3.2 Spatially and channel Squeeze-Excitation

Acoustic events are complex and diverse, so it is key to detect acoustic event categories and estimate acoustic source orientation to extract their directional features. The neural network model of attention mechanism pays more attention to effective information, ignores invalid information and extracts advanced features, so it is suitable for detecting and locating sound events.

In the baseline model CRNN, although the convolution kernel can be used to integrate the information of feature map spatial dimension and feature map dimension in the local receptive field, spatial and channel feature mapping cannot be independently learned. To solve this problem, the spatially and channel Squeeze-Excitation (scSE) module is added to the base model CRNN. The advanced feature diagram is obtained by compressing and weighting the channel characteristics. To enhance the channels that contribute more to classification and suppress the channels that contribute less to classification, the information on the localizations that play a key role in classification can be improved. The following describes the construction process of the scSE module.

Squeeze-and-Excitation network (SE) [25, 26] is the feature mapping of independent learning Spaces and channels. scSE is a variant of SE. It is a novel network architecture combined with the spatially Squeeze-Excitation (sSE) model and the channel Squeeze-Excitation (cSE) model. In this module, the channel and spatial relationship are considered at the same time, and the outputs of sSE and cSE are added and added to enhance the spatial coding ability of the convolutional layer network and improve the recognition effect of the neural network. The sSE model and cSE model are presented below.

In Fig. 5, the sSE model uses the convolution kernel size to realize the effects of channel excitation and spatial excitation, and the model introduces the attention

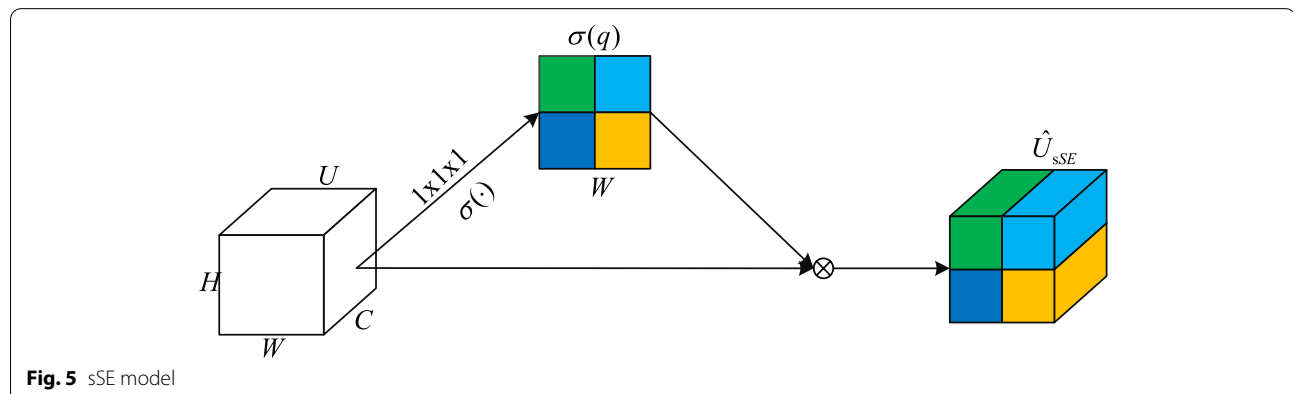


Fig. 5 sSE model

mechanism from the perspective of spatial relations. The featured graph of $H \times W \times C$ is subjected to a nonlinear operation of $1 \times 1 \times 1$ convolution dimension reduction and sigmoid function activation. After feature recalibration, output \hat{U} by multiplying the corresponding space of dimension U . Such as $W_{sq} \in R^{1 \times 1 \times C}$; output q through $q = W_{sq}U$, q is the feature tensor with channel number 1, and finally normalized to [0-1] through sigmoid. The operation function expression of this process is as follows:

$$\begin{aligned} \hat{U}_{sSE} &= F_{sSE}(U) \\ &= [\sigma(q_{1,1})u^{1,1}, \dots, \sigma(q_{i,j})u^{i,j}, \dots, \sigma(q_{H,W})u^{H,W}] \end{aligned} \quad (2)$$

where $\sigma(q_{i,j})$ represents the importance of (i, j) in the feature graph.

In Fig. 6, the cSE model uses the mutual stimulus between channels to build a feature mapping channel interdependence model. Inserting this module at a specific point in the network can get better results than the original advanced network, which has been well verified in the image classification task. Firstly, the unique feature map U of each channel was obtained by the global average pooling method, and the ReLU activation function was used to enhance the independence of each channel through two full connection layers with different weights, and the value was normalized to [0,1] through a sigmoid layer. That is, input $U=[u_1, u_2, \dots, u_C]$ into the channel, where the operation process of each channel is $u_i \in R^{H \times W}$, and the k value of each localization U output through the global pooling layer can be calculated by:

$$Z_k = \frac{1}{H \times W} \sum_i^H \sum_j^W u_k(i, j) \quad (3)$$

$$\hat{z} = W_1(\sigma(W_2z)), W_1 \in R^{c \times \frac{c}{p}}, W_2 \in R^{c \times \frac{c}{p}} \quad (4)$$

where \hat{Z} represents the importance of the feature of the i channel, p represents the ratio parameter, and W_1, W_2 represent the weight of two fully connected layers. The

independence of each channel is enhanced through the ReLU activation function and finally obtained $\sigma(\hat{z})$ through sigmoid normalization between 0 and 1. The process is calculated as follows:

$$\begin{aligned} \hat{U}_{cSE} &= F_{cSE}(U) \\ &= [\sigma(\hat{z}_1)u_1, \sigma(\hat{z}_2)u_2, \dots, \sigma(\hat{z}_c)u_c] \end{aligned} \quad (5)$$

Since sSE model takes spatial structure into account and cSE model takes channel arrangement into account, this paper establishes scSE model by adding and summing the outputs of the two models. The expression is as follows:

$$\hat{U}_{scSE} = \hat{U}_{sSE} + \hat{U}_{cSE} \quad (6)$$

The scSE module established in this paper is shown in Fig. 7.

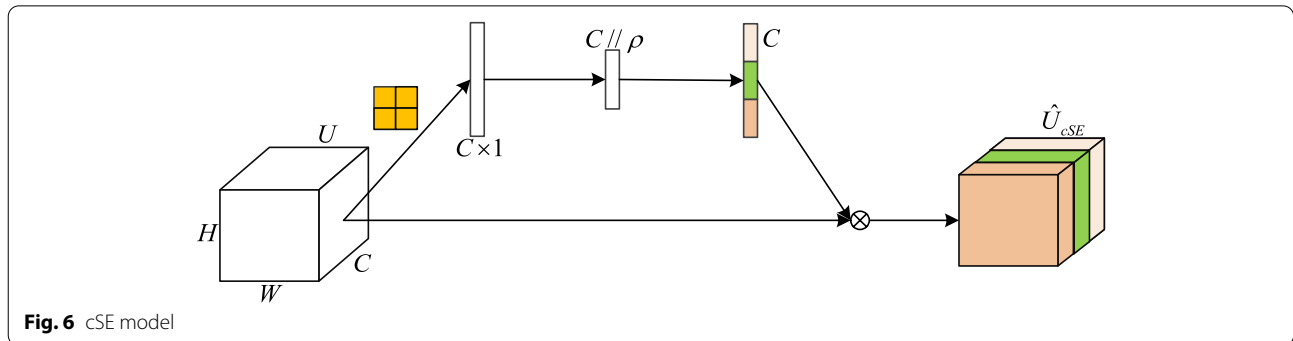
Add the scSE module shown in Fig. 7 to ResNet, which is added after the Exponential Linear Unit(ELU) activation function.

The scSE model combines the advantages of sSE and cSE to recalibrate feature maps from spatial and channel dimensions and combine the output information of the two modules to improve the detection effect of acoustic events.

3.3 Network model construction based on residual attention mechanism fusion

Considering the characteristics of ResNet structure, scSE, RNN and full connection layer, the Res-scSE-CRNN network structure proposed in this paper is shown in Fig. 8.

In Fig. 8, Res-scSE module is used to replace the convolution layer of CRNN network model to achieve the purpose and effect of high accuracy of SED and DOA. Two Bi-GRU layers were set up to obtain context information, and then the features extracted were dimensionally reduced through the full connection layer. Through the nonlinear operation of the sigmoid activation function and Tanh activation function, the categories of sound events and azimuth estimation were output respectively.



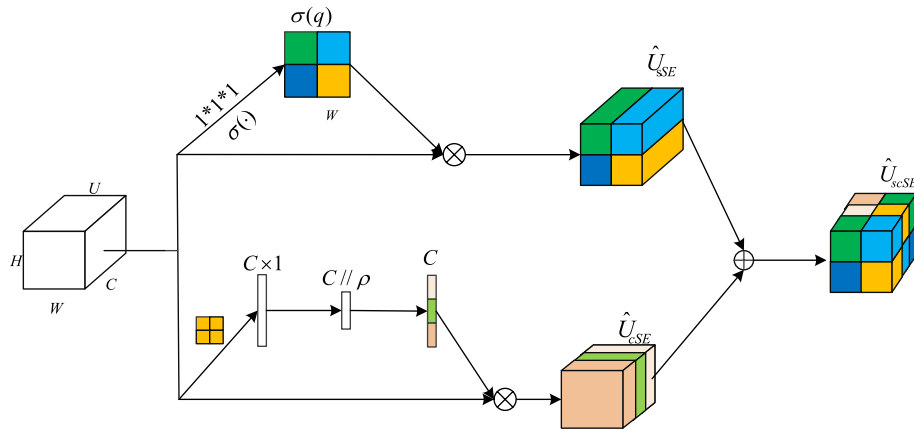


Fig. 7 scSE model

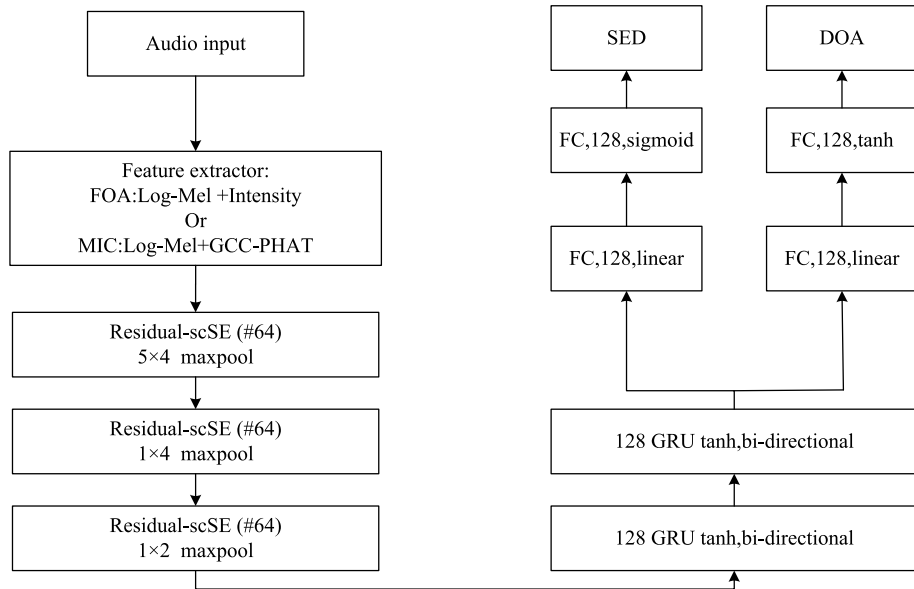


Fig. 8 Res-scSE-CRNN model

Log-mel harmonic strength vector and log-mel and GCC-PHAT features were used for input.

The improved Res-scSE module is shown in Fig. 9.

As you can see from Fig. 9, the module creates two jump connections before and after SE calibration. This double-jump connection allows the network to learn residual maps with and without SE recalibration simultaneously, and residual learning facilitates the training process by alleviating the gradient disappearance problem.

3.4 Evaluation indicators

Error rate (*ER*), F1-Score, Doa Error (*DE*), and frame recall rate (*FR*) are often used as indicators to evaluate the SEDL,

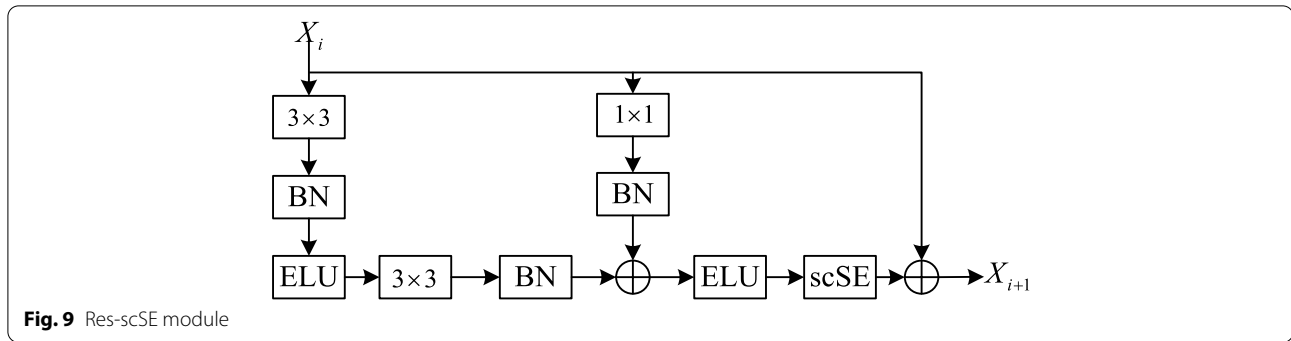
and these four indicators are used in this paper to evaluate the detection and localization effects of SEDL model.

(1) Evaluation index of detection

F1-Score and *ER* were used to evaluate the detection performance, as shown below:

The *F1*-Score is the harmonic average of accuracy *P* and recall. Where the accuracy calculation formula is:

$$P = \frac{\sum_{k=1}^K TP(k)}{\sum_{k=1}^K TP(k) + \sum_{k=1}^K FP(k)} \quad (7)$$



In the formula, TP is the true positive sample. In the k frame, the system predicted positive and actually positive; FP is a false positive sample, positively predicted by the system but negative in reality; FN represents a false negative sample, which is predicted by the system to be negative but actually positive.

The recall is expressed as:

$$R = \frac{\sum_{k=1}^K TP(k)}{\sum_{k=1}^K TP(k) + \sum_{k=1}^K FN(k)} \quad (8)$$

The $F1$ -Score is obtained by calculating P and R . The relationship between the two is:

$$F1 - score = \frac{2PR}{P + R} \quad (9)$$

$F1$ -Score can be calculated according to the following formula:

$$F1 - Score = \frac{2 \sum_{k=1}^K TP(k)}{2 \sum_{k=1}^K TP(k) + \sum_{k=1}^K FP(k) + \sum_{k=1}^K FN(k)} \quad (10)$$

The calculation expression of ER is:

$$ER = \frac{\sum_{k=1}^K S(k) + \sum_{k=1}^K D(k) + \sum_{k=1}^K I(k)}{\sum_{k=1}^K N(k)} \quad (11)$$

(2) Evaluation index of localizations

The localization performance is evaluated using DE and FR as follows:

The error rate of DOA is calculated as follows:

$$DE = \frac{1}{\sum_{k=1}^K D_E^k} \sum_{k=1}^K H(DOA_R^k, DOA_E^k) \quad (12)$$

where D_E^k represents DOA_R^k the total number of angles at the k moment, and $H(\cdot)$ represents the Hungarian algorithm to solve task allocation.

DOA frame recall rate, the calculation expression is:

$$FR = \frac{\sum_{k=1}^K 1(D_R^K = D_E^K)}{K} \quad (13)$$

where D_R^K refers to DOA_R^K the total number of angles of reference at the K th event. The calculated result is 1 when the condition is satisfied ($D_R^K = D_E^K$), and the result is obtained by adding this number over all the moments.

In an ideal environment, ER is close to 0, FR is close to 1, $F1$ -score is close to 1, and DE is close to 0, and the system performance is better.

4 Experimental results and analysis

4.1 Experimental environment and dataset

All experiments in this paper were carried out under the Framework of PyTorch. The experimental equipment adopts the processor Inter(R) Core(TM) i7-9700 CPU@3.00GHz, 32gb memory, 64-bit Windows 10 operating system, GPU GEFORCE RTX 2080 Ti, and GPU mode.

The main components of the SEDL model are feature extraction and M-CRNN. The training dataset is TAU Spatial Sound Events 2019-ambisonic and Microphone Array, Development dataset, which contains 400 1-minute recordings sampled at 48kHz and a total of 400 Sound fragments. And divided into four cross-validation fragments. The dataset provides data in two formats: the first is MIC format; The second kind of data is the first-order ambisonic format, namely FOA format.

Table 1 Layout table of microphone localizations

The microphone	ϕ	θ	r (cm)
M_1	45°	35°	4.2
M_2	−45°	−35°	4.2
M_3	135°	−35°	4.2
M_4	−135°	35°	4.2

Table 2 Dataset partitioning settings

The microphone	Division		
Dataset	Train	Validation	Test
Development	3,4	2	1

MIC format is as follows:

$$H_m(\phi_m, \theta_m, \phi, \theta, \omega) = \frac{1}{(\omega R/c)^2} \sum_{n=0}^{30} \frac{i^{n-1}}{h_n^{(2)}(\omega R/c)} (2n+1) p_n(\cos \gamma_m) \quad (14)$$

where m is the number of channels, (ϕ_m, θ_m) is the localization of azimuth and elevation angle of a specific microphone, $\omega = 2\pi f$ is the angular frequency, $R = 0.042m$ is the radius of the array, $c = 343m/s$ is the sound speed, and $\cos(\gamma_m)$ is the cosine angle between the microphone localization and the sound source. p_n is the non-normalized Legendre polynomial of degree N , and $h_n^{(2)}$ is the derivative of the parameters of the second spherical Hankel function. Table 1 shows the specific localizations of the microphone array.

4.2 Experimental results and analysis

Feature extraction stage: the size of a frame is set to 40 ms, the frameshift time is set to 20 ms, and the short-time Fourier points are set to 1024. In the frequency range of 0 to 22,500 Hz, the log-mel filter bank is set as 64. The microphone array size is 4. When log-mel spectrum and GCC-PHAT are selected as the feature input, the number of channels is 10 groups, and the dataset

format is MIC. When log-mel spectrum and Intensity Vector are used as input features, the input channels are 7 groups, and the dataset format is FOA. The division of the development dataset adopted in this paper is shown in Table 2.

The batch size for training was 32. The optimized method is Adaptive momentum(Adam). A total of 50 epochs were trained. After 30 epochs, the model converged to a relatively stable state.

4.2.1 Performance analysis of SEDL by M-CRNN

(1) Analysis of localization and detection effects of M-CRNN and other models

To analyze the generalization of the proposed model, ER , FR , $F1$ -Score, and DE were used to analyze Convolutional Neural Network (CNN) model, Convolutional Recurrent Neural Network (CRNN), some models from DCASE2019task3 challenge and Multiple-scale Convolutional Recurrent Neural Network (M-CRNN) model under the datasets of MIC and FOA formats. Make 200 predictions with Monte Carlo simulation to obtain results. The results are shown in Table 3.

It can be seen from the data results in Table 3 based on MIC format, compared with CNN, CRNN, some models from DCASE2019task3 challenge and M-CRNN, M-CRNN based on MIC format has the best ER and FR and the best detection performance, while DE index is not optimal and the localization performance is deficient. Compared with the baseline model CRNN, the model M-CRNN proposed in this chapter has a decrease of 0.24 in , an increase in $F1$ -Score and FR , however, the model experiment results in MIC dataset format show that M-CRNN model outperforms CRNN model. It can be seen from Table 3 that the experiment under FOA format dataset also performs better than the CRNN model in M-CRNN model and has a stronger model generalization ability. The corresponding evaluation indexes performed well in $F1$ -Score, ER and FR , and the detection performance was the best. The DE index was slightly higher, and the localization performance needed to be improved.

Table 3 Performance evaluation of each model based on the dataset in FOA and MIC format

The model name	On FOA format dataset				On MIC format dataset			
	DE (°)	FR (%)	ER	F1-score (%)	DE (°)	FR (%)	ER	F1-score (%)
CNN	19.9	81.3	0.38	75.3	19.8	75.3	0.31	81.2
CRNN	21.7	63.9	0.50	63.0	21.9	63.8	0.53	62.8
Chytas-UTH	18.6	82.4	0.29	75.6	19.8	81.2	0.31	75.3
FOA-baseline	24.6	85.4	0.28	85.7	28.5	79.9	0.34	85.4
M-CRNN	27.9	86.6	0.25	85.0	30.6	85.4	0.29	83.4

(2) Analysis the influence of hyperparameters of feature extraction on the detection and localization of M-CRNN model

To more intuitively analyze the influence of M-CRNN model on SEDL performance, the visual tools provided by DCASE competition were used to express the estimated information of SED and DOA in the audio segment in a graph, as shown in Fig. 10.

In Fig. 10, the abscissa represents time and the ordinate represents the phase. SED reference represents the reference value of SED and recorded the actual sound event category. SED predicted value represents the predicted sound event category. Azimuth reference refers to the reference value of azimuth angle and records the actual azimuth angle of the sound source. Azimuth predicted means the azimuth angle of sound source predicted by the model. Elevation reference refers to the reference value of elevation angle, which records the elevation angle of the actual sound source. Elevation reference refers to the elevation angle of the sound source predicted by the model.

As can be seen from Fig. 10, there are four overlapping segments in the detected audio. In terms of detection, according to the color comparison between SED Reference and SED predicted, the sound events in the overlapping part were detected normally with almost no error. In terms of localization, Azimuth in Azimuth Reference and Azimuth predicted have a deviation in phase. Elevation angle in elevation reference and elevation angle in

elevation predicted also have a deviation in phase. However, the overall predicted trend of azimuth and pitch in Fig. 10 is similar to the reference value.

4.2.2 Performance analysis of Res-scSE-CRNN model for SEDL

This part of the experiment first analyzes the optimal configuration of residual block and compressed excitation block in the Res-scSE-CRNN model, and discusses the detection and localization effects of the Res-scSE-CRNN model with or without the addition of attention block and different attention module states. The contribution of residual block and squeeze excitation block to SEDL in Res-scSE-CRNN model is studied, and the optimal value of the ratio parameter ρ is compared.

The data used in this section is the same as that in Section 4.2.1.

(1) Analysis the influence of the attention module on the performance of Res-scSE-CRNN model

The detection and localization effects of Res-scSE-CRNN model without the attention mechanism and when sSE, cSE, and scSE attention mechanisms are used are analyzed. Make 200 predictions with Monte Carlo simulation to obtain results, experimental results are shown in Table 4.

It can be seen from the data results in Table 4 based on FOA format, that the addition of modules based on the SE attention mechanism improves the detection and localization performance of the Res-scSE-CRNN model,

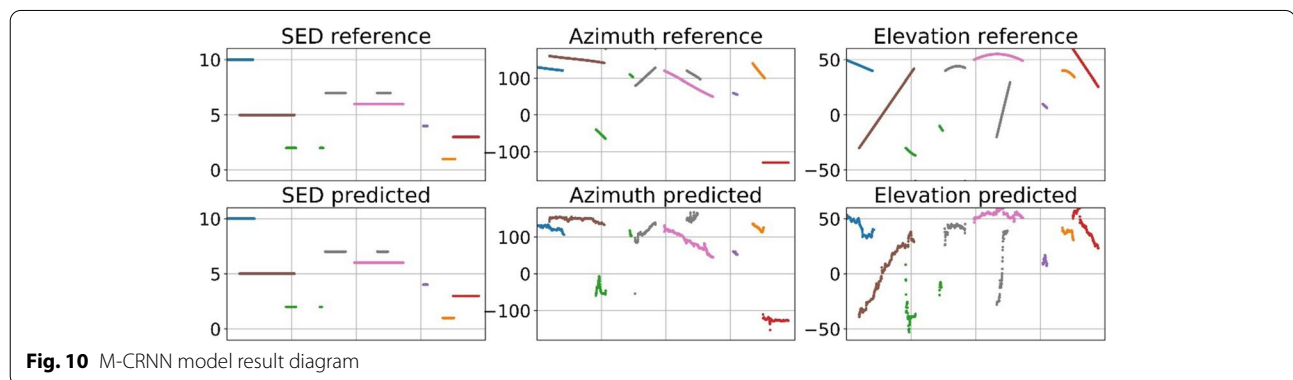


Table 4 Model performance comparison under different attention mechanisms based on FOA and MIC format dataset

The model name	On FOA format dataset				On MIC format dataset			
	DE (°)	FR (%)	ER	F1-score (%)	DE (°)	FR (%)	ER	F1-score (%)
sSE	7.6	88.4	0.24	87.8	8.0	87.4	0.26	86.8
cSE	7.5	89.3	0.25	87.7	7.9	88.3	0.28	86.7
scSE	5.1	89.6	0.21	88.4	5.5	88.5	0.23	87.4
Not using attention	13.5	84.7	0.28	82.1	14.1	83.4	0.31	83.1

while the evaluation indexes of the Res-scSE-CRNN model without the addition of the attention mechanism are all reduced. For example, the addition of sSE resulted in a 0.03 decrease in *ER*, a 5.7% increase in *F1-Score*, a 5.9° decrease in *DE*, and a 3.7% increase in *FR* compared with the addition of the attention module. The cSE showed a 0.04 decrease in *ER*, a 5.6% increase in *F1-Score*, a 6.0° decrease in *DE*, and a 3.6% increase in *FR* compared to those without the attention module. Therefore, using spatial attention or channel attention can improve the performance of the model. Comparing sSE and cSE, the contribution of the two kinds of attention is similar. The scSE obtained by fusing two different attention modules improved significantly in all four indicators.

It can be seen from the data results in Table 4 based on MIC format, compared with the model performance of different attention mechanisms for the dataset based on the MIC format, all the overall indicators decline, but the decline range is not significant.

(2) Analysis the influence of ratio parameters on the performance of Res-scSE-CRNN model

The best model “Res-scSE-CRNN of Conv-Standard-Post configuration module” obtained in Experiment 1 of this section was used for the comparison experiment. Experimental parameter setting: Set other hyperparameters as fixed values, which are the ratio parameters in cSE module. Change the value of ratio parameters to observe the influence of ratio parameters on the Res-scSE-CRNN model. Make 200 predictions with Monte Carlo simulation to obtain results, experimental results are shown in Table 5.

It can be seen from the data results in Table 5 based on FOA format, when the ratio parameter of the attention mechanism is 16, the best effect of *DE* is 4.7; when it is 1, the best effect *FR* is 90.1%; when it is 4, *ER* and *F1-Score* are the best, 0.21 and 88.4%, respectively. Overall analysis, two optimal indexes of the four indexes appear when it is 4, and the difference between *DE* and the optimal value is 0.2°, and the difference between *FR* and the optimal value is a drop of 0.5 percentage points. The results show that the best performance is achieved when the ratio parameter ρ of the attention mechanism is 4.

It can be seen from the data results in Table 5 based on MIC format, there is little difference between the model performance comparison under different ratio parameters ρ based on MIC dataset and that under different ratio parameters based on FOA dataset.

(3) Comparison of detection and localization performance between the Res-scSE-CRNN model proposed in this paper and other models

ER, *FR*, *F1-Score*, and *DE* evaluation indexes were used to analyze the SEDL effects of the CNN, CRNN, M-CRNN models, some models from DCASE2019task3 challenge and Res-scSE-CRNN model proposed in this paper. Make 200 predictions with Monte Carlo simulation to obtain results, experimental results are shown in Table 6.

It can be seen from the data results in Table 6 based on FOA format, compared with CNN, CRNN, M-CRNN, some models from DCASE2019task3 challenge and other models based on FOA dataset, Res-scSE-CRNN model has the best four indicators. In terms of detection, *ER* of Res-scSE-CRNN model decreases by 0.04 compared with M-CRNN model. *FR* increased by 4.2%, *ER* decreased by 0.07 and *F1-Score* increased by 2.7% compared with FOA-baseline model. Compared with M-CRNN model, *DE* decreased by 22.8° and *FR* increased by 3.0%.

From the analysis of the overall performance of the model, it can be found that compared with M-CRNN model, the Res-scSE-CRNN model is significantly improved in localization and detection.

It can be seen from the data results in Table 6 based on MIC format, the performance of different models based on MIC datasets decreases compared with that based on FOA datasets.

All the above experimental data are Make 200 predictions with Monte Carlo simulation to obtain results, which are of reference value.

To more intuitively observe the influence of Res-scSE-CRNN model on SEDL performance, visual tools were selected to draw, and the detection and localization effects were drawn as shown in Fig. 11.

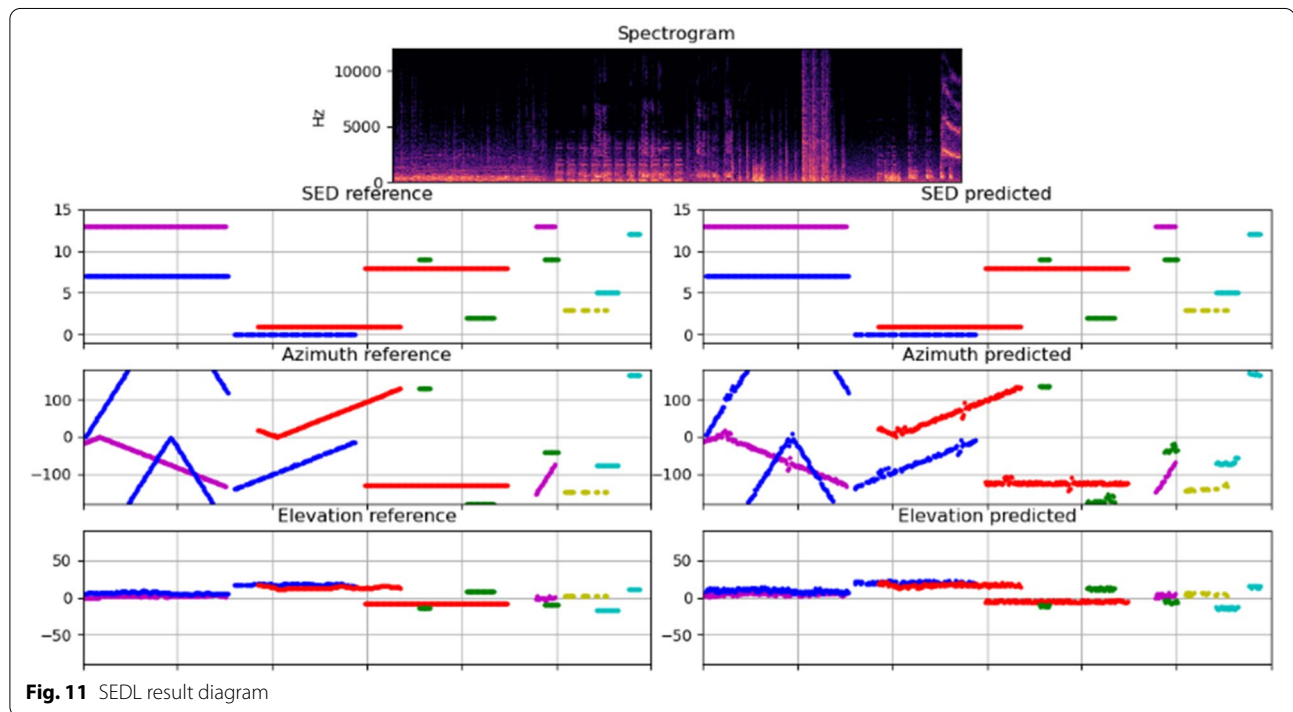
As can be seen from Fig. 11, 8 different sound events and their related azimuth and elevation were detected in

Table 5 Model performance comparison under different attention mechanisms based on FOA and MIC format dataset

ρ	On FOA format dataset				On MIC format dataset			
	DE (°)	FR (%)	ER	F1-score (%)	DE (°)	FR (%)	ER	F1-score (%)
1	4.9	90.1	0.22	87.9	5.0	89.1	0.25	86.7
2	4.8	89.3	0.23	88.3	4.9	88.3	0.26	87.4
4	4.9	89.6	0.21	88.4	5.3	88.6	0.23	87.3
8	5.0	88.9	0.22	85.6	5.1	87.9	0.25	84.5
16	4.7	89.0	0.23	86.7	4.9	88.0	0.24	85.6

Table 6 Performance comparison of each model on FOA and MIC format dataset

The model name	On FOA format dataset				On MIC format dataset			
	DE (°)	FR (%)	ER	F1-score (%)	DE (°)	FR (%)	ER	F1-score (%)
CNN	19.9	81.3	0.38	75.3	19.8	75.3	0.31	81.2
CRNN	21.7	63.9	0.50	63.0	21.9	63.8	0.53	62.8
Chytas-UTH	18.6	82.4	0.29	75.6	19.8	81.2	0.31	75.3
FOA-baseline	24.6	85.4	0.28	85.7	28.5	79.9	0.34	85.4
M-CRNN	27.9	86.6	0.25	85.0	30.6	85.4	0.29	83.4
Res-scSE-CRNN	5.1	89.6	0.21	88.4	5.3	88.6	0.23	87.3



this audio clip. In terms of detection, compared with the reference value of SED, the predicted value of SED in the Fig. 11 was accurately identified. In terms of localization, the deviation between the predicted value of Azimuth and the reference value of Azimuth and the predicted value of Elevation in Fig. 11 is relatively low.

5 Conclusions

Aiming at the insufficient feature extraction of CRNN model at a single scale, a sound event detection and localization method based on M-CRNN was established based on M-CNN structure of multi-scale convolution fusion combined with RNN. The results show that the *ER*, *F1-Score*, *DE*, and *FR* of M-CRNN network model are 0.25, 85.0%, 27.9%, and 86.6% on FOA dataset. In

MIC dataset, *ER*, *F1-Score*, *DE*, and *FR* are 0.29, 83.4%, 30.6%, and 85.4%, respectively. Compared with other SEDL models, the SE models are superior to most other acoustic event detection and localization models. It can be seen that this method has some advantages in SEDL research, but its localization performance is insufficient.

Although the detection effect of M-CRNN proposed in this paper is good, the localization effect still needs to be improved on the basis of enhancing the expression of key channel features and spatial features and avoiding network degradation. Therefore, a Res-scSE-CRNN sound event detection and localization method based on residual fusion attention mechanism is proposed. The results show that the *ER*, *F1-Score*, *DE*, and *FR* of Res-scSE-CRNN model on FOA dataset are 0.21,

88.4%, 5.1°, and 89.6%, respectively. In MIC dataset, *ER*, *F1*-Score, *DE*, and *FR* were 0.23, 88.4%, 5.6°, and 88.6%, respectively. Compared with other SEDL models, Res-scSE-CRNN model showed significant improvement in detection and localization effect, especially in localization performance.

Acknowledgements

This work was supported by the natural science foundation of Heilongjiang Province (No. LH2020F033) and the national natural science youth foundation of china (No.11804068).

Authors' contributions

Chaofeng Lan contributed to the conception of the study and contributed significantly to analysis and manuscript preparation; Lei Zhang made important contributions in making adjustments to the structure, revising the paper, english editing and major revisions of this manuscript; Yuanyuan Zhang performed the experiment, the data analyses and wrote the manuscript; Lirong Fu made significant contributions to the structure, English modification and data analysis of the paper during the major revision process; Dr. Meng Zhang made significant contributions to the structure, English modification in the second round manuscript preparation. Dr. Meng Zhang also conducted the final manuscript review and revision; Chao Sun and Yulan Han made important contributions in making adjustments to the proofread English. All authors reviewed the manuscript.

Funding

This research received by the natural science foundation of Heilongjiang Province (No. LH2020F033) and the national natural science youth foundation of china (No.11804068).

Availability of data and materials

All data included in this study are available upon request by contact with the corresponding author to the following link: <http://cetong.hrbust.edu.cn/2021/0818/c3451a64186/page.htm>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Springer Nature remains neutral about concerning jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of School of Measurement and Communication Engineering, Harbin University of Science and Technology, 150080 Harbin, People's Republic of China. ²Beidahuang Industry Group General Hospital, 150088 Harbin, People's Republic of China. ³Mechanical and Electrical Engineering College, Hainan University, 570228 Haikou, People's Republic of China. ⁴School of Electronics and Communication Engineering, Guangzhou University, 510006 Guangzhou, People's Republic of China.

Received: 19 March 2022 Accepted: 16 November 2022

Published online: 05 December 2022

References

1. L. Weijie, L. Bo, Modern Electronic Technique. **42**(12), 45–47 (2022)
2. X. Zhou, X. Zhuang, M. Liu et al., *HMM-based Acoustic Event Detection with AdaBoost feature selection* (Springer, Berlin, Heidelberg, 2007), pp.345–353
3. W. Junqin, W. Yingfu, Speech Separation Based on GCC-NMF. *J. Jiangxi Univ. Sci. Technol.* **41**(05), 65–72 (2020)

4. H. Phan, M. Maaß, R. Mazur et al., Random Regression Forests for Acoustic Event Detection and Classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(1), 20–31 (2014)
5. K.J. Piczak, *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Environmental Sound Classification with Convolutional Neural Networks (IEEE, Boston, 2015), pp. 1–6
6. T.H. Vu, J.C. Wang, Acoustic Scene and Event Recognition Using Recurrent Neural Networks. *Detect. Classif. Acoust. Scenes Events.* **1**, 10–15 (2016)
7. S. Jianan, Design of binary network system for one-dimensional time series signal (Beijing University of Posts and Telecommunications, Beijing, 2020), pp.10–15
8. R.E. Caki, G. Parascandolo, T. Heittola et al., Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(6), 1291–1303 (2017)
9. S. Adavanne, A. Politis, T. Virtanen, *2018 International Joint Conference on Neural Networks (IJCNN)*. Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features (IEEE, Rio de Janeiro, 2018), pp. 1–7
10. Q. Kong, Y. Xu, I. Sobieraj, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4. IEEE/ACM Transactions on Sound Event Detection and Time-frequency Segmentation from Weakly Labelled Data (2019), pp. 777–787. <https://doi.org/10.1109/TASLP.2019.2895254>
11. T. Iqbal, Y. Xu, Q. Kong, et al., *2018 26th European Signal Processing Conference (EUSIPCO)*. Capsule routing for sound event detection (IEEE, Rome, 2018), pp. 2255–2259
12. L. Yang, J. Hao, X. Gu, Z. Hou, Audio Label consistency constraint CRNN Sound Event Detection. *J. Electron. Inf. Technol.* **44**(03), 1102–1110 (2022)
13. Y. Huang. Noise cases, sound detection, classification and localization (University of electronic science and technology, 2021), <https://doi.org/10.27005/d.cnki.gdzku.2021.003635>.
14. J. Lee, J. Nam, Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE Sig. Process. Lett.* **24**(8), 1208–1212 (2017)
15. C. Xinxing, *Research on multi-scale feature fusion and Data augmentation method for sound scene classification* (Chongqing University, Chongqing, 2019), pp.8–10
16. Z. Weizhe, QIU Peng, Wei Juan. Voice Recognition and Detection Based on Multi-scale Attention Fusion in Weak Label Environment. *Comput. Sci.* **47**(05), 120–123 (2020)
17. H. Xinyuan. Research on deep network model for acoustic event localization and detection (Yanshan University, 2021), <https://doi.org/10.27440/d.cnki.gysdu.2021.001846>.
18. I. Trowitzsch, C. Schymura, D. Kolossa et al., Joining sound event detection and localization through spatial segregation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 487–502 (2019)
19. Y. Cao, Q. Kong, T. Iqbal, et al., Polyphonic Sound Event Detection and Localization using a two-stage Strategy. (2019). ArXiv Preprint [arXiv:1905.00268](https://arxiv.org/abs/1905.00268)
20. R. Ranjan, S. Jayabalan, T.N.T. Nguyen, et al., Sound Event Detection and Direction of Arrival Estimation using Residual Net and Recurrent Neural Networks. (2019). ArXiv preprint [arXiv:1902.00260](https://arxiv.org/abs/1902.00260)
21. E.L. Tan, R. Ranjan, S. Jayabalan, Sound Event Detection and Localization using ResNet RNN and Time-delay DOA. *Detect. Classif. Acoust. Scenes Events Chall.* **26**(7), 1751–1760 (2019)
22. J. Naranjo-Alcazar, S. Perez-Castanos, J. Ferrandis, et al., Sound Event Localization and Detection using Squeeze-Excitation Residual CNNs. (2020). arXiv preprint [arXiv:2006.14436](https://arxiv.org/abs/2006.14436)
23. L. Min, M. Zhenjiang, X. Wanru, A CRNN-based attention-seq2seq model with fusion feature for automatic Labanotation generation. *Neurocomputing.* **454**, 430–440 (2021)
24. L. Bin, Z. Junyue, C. Jie. Efficient Residual Neural Network for Semantic Segmentation. *Pattern Recognit. Image Anal.* **31** (2), (2022)
25. W. Zhanguo, S. Yaping, L. Ya, Improved IMAGE segmentation algorithm of logistics tray based on squeeze excitation cavity convolution in U-NET Network. *J. Packag.* **13**(05), 35–41 (2021)
26. J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 2011–2023 (2020)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.