

Article

Environmental Sound Classification Framework Based on L-mHP Features and SE-ResNet50 Network Model

Mengxiang Huang ¹, Mei Wang ^{1,2,*}, Xin Liu ^{1,2,*}, Ruixiang Kan ³ and Hongbing Qiu ^{3,4}

¹ College of Information Science and Engineering, Guilin University of Technology, Guilin 541006, China; 1020210969@glut.edu.cn

² Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541006, China

³ School of Information and Communication, Guilin University of Electronic Technology, Guilin 541006, China; bbklasnic@glut.edu.cn (R.K.); qiuhb@guet.edu.cn (H.Q.)

⁴ Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin University of Electronic Technology, Guilin 541006, China

* Correspondence: mwang@glut.edu.cn (M.W.); liuxin2017125@glut.edu.cn (X.L.)

Abstract: Environmental sound classification (ESC) tasks are attracting more and more attention. Due to the complexity of the scene and personnel mobility, there are some difficulties in understanding and generating environmental sound models for ESC tasks. To address these key issues, this paper proposes an audio classification framework based on L-mHP features and the SE-ResNet50 model and improves a dual-channel data enhancement scheme based on a symmetric structure for model training. Firstly, this paper proposes the L-mHP feature to characterize environmental sound. The L-mHP feature is a three-channel feature consisting of a Log-Mel spectrogram, a harmonic spectrogram, and a percussive spectrogram. The harmonic spectrogram and percussive spectrogram can be obtained by harmonic percussive source separation (HPSS) of a Log-Mel spectrogram. Then, an improved audio classification model SE-ResNet50 is proposed based on the ResNet-50 model. In this paper, a dual-channel data enhancement scheme based on a symmetric structure is promoted, which not only makes the audio variants more diversified, but also makes the model focus on learning the time–frequency mode in the acoustic features during the training process, so as to improve the generalization performance of the model. Finally, the audio classification experiment of the framework is carried out on public datasets. An experimental accuracy of 94.92%, 99.67%, and 90.75% was obtained on ESC-50, ESC-10 and UrbanSound8K datasets, respectively. In order to simulate the classification performance of the framework in the actual environment, the framework was also evaluated on a self-made sound dataset with different signal-to-noise ratios. The experimental results show that the proposed audio classification framework has good robustness and feasibility.

Keywords: acoustic image analysis; time–frequency analysis; data enhancement; convolutional neural network



Citation: Huang, M.; Wang, M.; Liu, X.; Kan, R.; Qiu, H. Environmental Sound Classification Framework Based on L-mHP Features and SE-ResNet50 Network Model. *Symmetry* **2023**, *15*, 1045. <https://doi.org/10.3390/sym15051045>

Academic Editor: José Carlos R. Alcántud

Received: 17 April 2023

Revised: 5 May 2023

Accepted: 6 May 2023

Published: 9 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Sound is an essential way for people to accept or transmit information, and it is often used in our daily work to carry information. More and more research fields and projects are attracting acoustic research and development personnel, and automatic speech recognition may be the most representative task in the field of sound. Many challenges remain in the sound domain [1], one of which is Environmental Sound Classification (ESC). Environmental sounds contain many kinds of sounds that exist in our daily lives, including but not limited to dog barking, knocking on the door, baby crying, etc. The purpose of ESC is to make computers identify and classify these sounds correctly, which has a profound impact on people's daily lives. For example, relevant departments are aware of the timeliness to emergency situations; installing acoustic monitoring and analyzing

abnormal sounds in the streets enables a faster response, e.g., cars can avoid an accident by recognizing a child screaming in a location near the car and slowing down. The relevant technology in the field of ESC is of critical importance in noise pollution analysis, environmental monitoring, the development of smart cities and many other applications. This paper will focus on several key issues in ESC tasks, such as the ability to characterize acoustic features, accuracy of classification, and the generalization of performance of the model and so on. Related works on the above issues are described in the next section.

1.1. Related Works

Different from speech, music, and other sound signals, the sound signal of the environment has complex characteristics. Meanwhile, it is often mixed with complex and flexible noises in real-world situations. Therefore, how to characterize environmental sound is gradually becoming a major problem in this field. The acoustic features commonly used in audio recognition tasks, such as Spectral Contrast [2], Mel spectrogram, Mel-scale Frequency Cepstral Coefficients [3] (MFCC), etc., are proposed for speech or musical sound recognition, so the ability to characterize environmental sounds is limited. In order to solve this problem, many scholars adopt to fuse features to jointly characterize environmental sounds [4,5]. For example, reference [6] analyzed a variety of acoustic features and demonstrated their utility through experiments. The experimental results and analysis show that the MGCC feature fused with MFCC and GFCC (Gammatone frequency cepstral coefficients) [7] has the best ability to characterize environmental sounds. In [8], Xu et al. proposed to use three hamming windows of different lengths to extract short, middle, and long-term features of a sound signal, and sent them to a customized three-stream Convolutional Neural Network (CNN) for training, achieving great classification accuracy for Frog datasets. In [9], Guzhov A et al. adopted a three-channel feature that decided to separate the Log-STFT spectrogram along its frequency domain into three parts which were mapped onto three input channels, respectively, and achieved good classification results. Nasiri A et al. [10] adopted the Log-Mel spectrogram to replicate three channels as input features, combined it with the ResNet-50 model as a classification model, and achieved good experimental accuracy. However, it is difficult to directly use its feature extraction scheme to cope with the actual demands in a real-world situation. Therefore, it is necessary to introduce the solutions of feature engineering theory or methods on this basis.

In addition to the selection of acoustic features, the selection of audio classification models is also an important issue in this field. Previously, the general trend in ESC tasks was to design audio-domain-specific architectures. Many audio classification models are implemented based on the CNN which is one of the most commonly used network structures in classification tasks [11–13]. Piczak et al. [14] applied the CNN to a sound classification task for the first time and showed good performance on public datasets by learning Log-Mel spectrogram features. In addition, reference [5] proposed a double two-stream CNN, which was combined with a feature fusion scheme composed of four acoustic features as input, and achieved fabulous accuracy on the UrbanSound8K dataset. B. Bahmei et al. [15] proposed using CNN models as a feature extractor and RNN as a classifier, because the RNN model has the ability to combine the information in the temporal domain, and obtained a tremendous classification effect.

In recent years, more and more popular and adaptive network models in image fields have been redesigned and applied in ESC [16,17] due to the development of the transfer learning method. These studies used deep full convolution models. Surprisingly, these models were pre-trained on ImageNet and fine-tuned using the spectrogram representation of audio samples. Since the spectrogram of audio events shows some local correlation similar to that of images, the pre-trained model on large image datasets also performs well in audio classification. Therefore, introducing the pre-trained classification model to ESC tasks can accelerate the convergence speed of the model and improve the classification performance to a certain extent. For example, in reference [10], the ResNet-50 model pre-trained by the ImageNet dataset is applied to ESC tasks, and good results are obtained.

In reference [17], the DenseNet-201 model is used as an audio classification model, and the Log-Mel spectrogram is copied into three channels as input features. However, up to now, it is still a major difficulty in this field to propose an audio classification model with excellent performance and strong generalization ability, and to match it with a feature fusion strategy that can fully characterize environmental sound signals.

The environmental sound dataset is different from the image field with its extremely rich data samples. Although there are new sound datasets being released, the smaller sample size is still not conducive to the training and learning of a network model. For this problem, Piczak [14] proposed a set of audio data enhancement schemes, including random combinations of time shifting, pitch shifting, and time stretching, and his experiments show that this scheme can obtain satisfactory classification results on the UrbanSound8K dataset. In reference [18], J. Salamon et al. proposed five sound data enhancement schemes and applied them to model training, which reduced the impact of small-capacity sound datasets on classification performance to a certain extent. In addition to the above traditional data enhancement methods, reference [15] proposed an offline data enhancement scheme based on deep learning; they used Deep Convolutional Generative Adversarial Networks (DCGANs) [19] to learn the Log-Mel spectrogram of an original audio, and then output images similar to the original Log-Mel spectrogram for an extended dataset, and achieved great results. According to the above studies, the data enhancement technology carries out a small scale transformation of original samples to generate similar sample variants, which can extend a dataset and alleviate the impact of a small dataset on a model performance to a certain extent. Therefore, it is necessary to set up an effective data enhancement scheme at the data preprocessing stage.

Although many scholars have put forward their own solutions to the above problems, there are still many difficulties to be solved in ESC tasks. For example, when the audio classification frameworks proposed in some papers are used for verification experiments, we find that once a small amount of noise is added to the dataset, the classification accuracy of the model will decrease by 5–15% [9–17], indicating that the robustness of these models is unstable. So, it is necessary to put forward novel solutions and improvement measures on the basis of existing research schemes.

1.2. Contributions

In order to solve the difficulties mentioned above, this paper proposes an audio classification framework for ESC tasks which is focused on the L-mHP acoustic feature and SE-ResNet50 model. Different from other schemes based on spectrogram-based features and the ResNet-50 network model [9,10,17,18], a novel combined feature scheme, L-mHP, is designed and the SE-ResNet50 model is improved for audio classification. The L-mHP feature consists of three kinds of spectrograms, which complement each other to represent environmental sounds better. The Squeeze-and-Excitation (SE) attention block is embedded in each Stage layer in the SE-ResNet50 model to improve the model's ability to extract critical information. Meanwhile, this study also improves a set of dual-channel data enhancement schemes based on a symmetric structure for model training. In this paper, the proposed audio classification framework and dual-channel data enhancement scheme are applied to public sound datasets to evaluate their classification performance. For simulating the real environment, we make our own sound datasets by adding noise with different signal-to-noise ratios, and conduct verification experiments on the dataset, so as to explore the robustness of the proposed audio classification framework and the feasibility of application in the actual environment.

In brief, our main contributions in this paper can be summarized as follows:

- (1) A novel three-channel acoustic feature combination strategy of L-mHP to characterize environmental sound is proposed. Each channel of the feature focuses on different spectral characteristics, which means each channel describes the environmental sound signal from a different dimension and complements each other.

- (2) An improved audio classification model SE-ResNet50 is energized, which is designed based on the ResNet-50 model. By introducing the SE attention mechanism, the model is expected to pay more attention to the more useful feature channels for the classification task during the training process. It is also expected to enhance the ability of the model to extract key information.
- (3) A set of dual-channel data enhancement schemes based on a symmetric structure is improved and implemented. In the process of feature extraction, this enhancement scheme is used to transform the original audio and spectrogram to generate more diversified variants, which helps the classification model learn more about the time–frequency mode of acoustic features during training and improves the performance of the model.

2. Proposed Methods

This section mainly introduces the limitations of the original scheme, the improvement measures proposed in this paper, and illustrates the principle, technical scheme and function of these improvement measures in detail.

2.1. L-mHP Feature Extraction Scheme

There are always environmental sound signals containing various noises, which are quite different from speech, music, and other sound signals. So it is difficult to extract useful information from an original sound signal. To solve this problem, common practice is to map the original audio data to the Log-Mel spectrogram or MFCC spectrogram [3,20]. The Log-Mel spectrogram is a time–frequency feature designed to imitate the auditory system of the human ear. It is obtained by performing Short Time Fourier Transform (STFT) [21] on the original sound signal first, then sending it to the Mel filter bank to map its frequency to the Mel frequency and take the logarithm. The SE-ResNet50 audio classification model proposed in this paper is implemented by introducing the SE attention mechanism based on the ResNet-50 model, which is a commonly used classification model in the visual domain [22]. ResNet-50 is also used as the baseline model set in this experiment. Like other visual domain classification models, its input features are three channels (corresponding to red, green, and blue in the image), and the Log-Mel spectrogram we extracted is a single-channel spectrogram. In order to solve this problem, many scholars adopt the scheme of copying the Log-Mel spectrogram to form a three-channel feature [10] or matching the Log-Mel spectrogram with two zero-value channels. The main disadvantage of this approach is that it will cause unnecessary information redundancy or loss of key information [9].

In order to make up for the shortcomings, this paper explores a novel acoustic feature combination L-mHP, which can be applied to the visual domain classification model. Specifically, L-mHP is a three-channel feature composed of the Log-Mel spectrogram, harmonic spectrogram, and percussive spectrogram. The harmonic spectrogram and the percussive spectrogram are obtained by harmonic percussive source separation (HPSS) [23] of the Log-Mel spectrogram. Each channel of this feature focuses on different spectral characteristics of sound and are complementary to each other. The Log-Mel spectrogram is responsible for capturing the global characteristics of sound signals under this circumstance. The harmonic spectrogram focuses on describing the frequency distribution and frequency band activity of the sound. Meanwhile, the percussive spectrogram has obvious vertical structure, and it focuses on reflecting the frequency variation trend between consecutive frames of the sound signal. Figure 1 shows the extraction process of the L-mHP features.

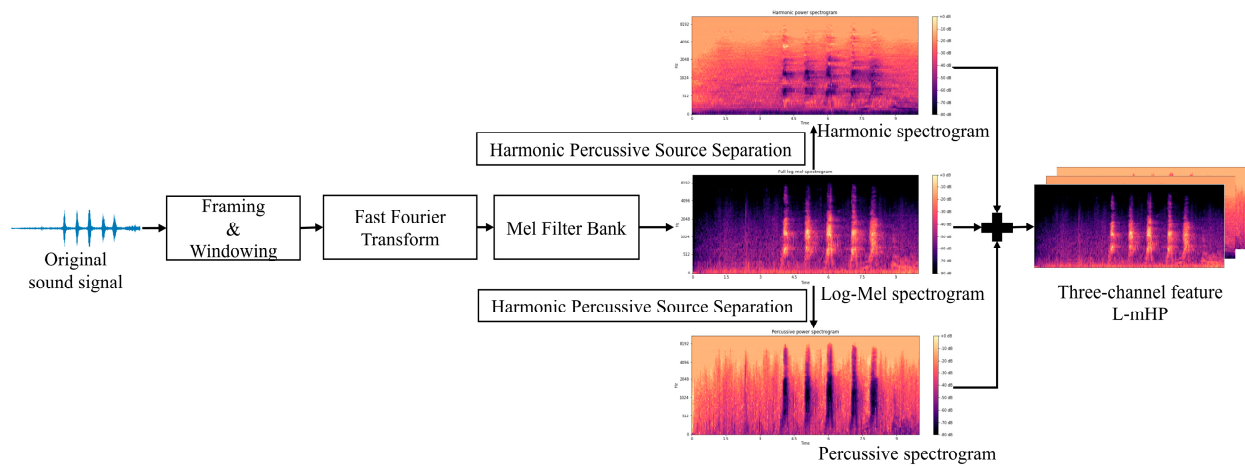


Figure 1. The extraction process of L-mHP features.

- (1) **Framing and Windowing:** Mapping the audio signal from the time domain to the frequency domain requires a Fourier transform of the audio signal. However, Fourier transform cannot be directly applied to a non-stationary sound signal, so the audio signal needs to be divided into frames. The n -th frame audio is recorded as $s(n)$, the frame length is set to 1024, and the frame shift is set to 512. In order to reduce the spectrogram leakage caused by framing, it is necessary to add a hamming window $\omega(n)$ to the frame signal $s(n)$:

$$\omega(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1 \quad (1)$$

where N is the total number of sampling points.

- (2) **Fast Fourier Transform (FFT):** The complex spectrogram of each frame of audio is obtained by Fast Fourier Transform. The input signal is $s(n)$, and the Discrete Fourier Transform (DFT) formula of the signal is:

$$S(k) = \sum_{n=0}^{N-1} s(n) * \exp\left(\frac{-j2\pi kn}{N}\right), 0 \leq k \leq N-1, \quad (2)$$

where N represents the number of sampling points for DFT transformation and $S(k)$ represents the value of the k -th frequency point. Then it is supposed to take the modulus square of the resulting complex spectrogram to get the power spectrogram. In practice, the STFT [21] is used instead of FFT. The STFT is an improved method based on steps 1 and 2. The audio signals are executed for framing, windowing, and FFT automatically by STFT with specified window length.

- (3) **Mel Filter Bank:** The power spectrogram is mapped to the Mel scale through the Mel filter bank [14]. The conversion formula consists of Mel frequency $Mel(f)$ and frequency f is:

$$Mel(f) = \frac{1000}{\log 2} \log\left(\frac{f}{1000} + 1\right). \quad (3)$$

Then the Log-Mel spectrogram is obtained by logarithmic function and the Mel-scale power spectrogram.

- (4) **HPSS:** HPSS [24] is performed on the spectrogram to separate the mixed signal in the spectrogram into harmonic components and percussive components. The separation process requires the median filter, whose calculation formula is:

$$\gamma(a) = median\{x(a-j, a+j), j = (l-1)/2\} \quad (4)$$

where, $x(a)$ is the input signal sequence, l is the filter length, and $median\{\}$ is the median filter function. The absolute values of the time domain component $S(t)$ and the frequency domain component $S(f)$ of the spectrogram signal S are input into Equation (4). Then the output is the harmonic enhancement fraction H and the percussive enhancement fraction P of the spectrogram signal S :

$$H = median\{|S(f)|, l_H\}$$

and

$$P = median\{|S(t)|, l_P\}$$
(5)

where l_P and l_H correspond to the filter length of the percussive enhancement part and the filter length of the harmonic enhancement part, respectively. The relative components M_H and M_P in the spectrogram are calculated by using the harmonic enhancement fraction and the percussive enhancement fraction based on the spectrogram.

$$M_H = \frac{H^P}{H^P + P^P}$$

and

$$M_P = \frac{P^P}{H^P + P^P},$$
(6)

where P is the energy coefficient. The two relative components are combined with the original spectrogram signal to obtain the spectrogram harmonic component and the spectrogram percussive component.

Figure 2a,b show the L-mHP acoustic feature of a glass breaking sound and a siren sound, respectively. It can be seen from Figure 2a,b that each channel of the L-mHP characteristic represents the spectral characteristics. The Log-Mel spectrogram contains the global characteristics of the sound. Meanwhile the harmonic spectrogram and the percussive spectrogram supplement the sound signal along the frequency axis and the time axis on the basis of the Log-Mel spectrogram respectively. Therefore, it is theoretically feasible and effective to use the L-mHP feature as the representation feature of the sound signal. Subsequently, this paper carried out relevant simulation experiments on the baseline model. The specific experimental results are shown in Table 1, which proves that this acoustic feature can effectively characterize the environmental sound and improve the experimental performance.

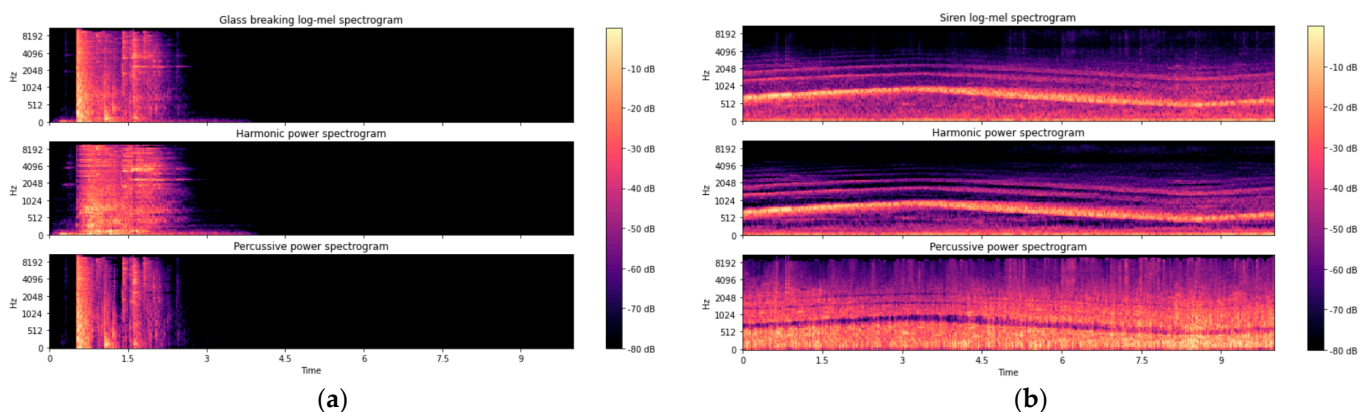


Figure 2. The L–mHP feature of different environmental sound. (a) Glass Breaking. (b) Siren.

Table 1. Experimental performance of improved dual-channel data enhancement scheme.

Data Enhancement Scheme	Accuracy (%)	Processing a Single Audio (ms)
Original Scheme [10]	90.2	17.3
Improved Scheme	91.31	30.7

2.2. Improved Dual-Channel Data Enhancement Scheme Based on Symmetric Structure

Since the environmental sound datasets are different from the image field with rich sample data, small-capacity sound datasets are prone to model overfitting and other problems during the training process, which affects the classification accuracy of the model. Therefore, it is necessary to apply data enhancement technology to the training process of a model. The novel solutions proposed in our method will be of great benefit and reduce the influence of the small-capacity datasets on model performance to some extent and improve the performance of the model. In this study, a dual-channel data enhancement scheme based on a symmetric structure is improved on the basis of [10]. The scheme consists of two data enhancement channels, which perform data enhancement processing on the original audio and spectrogram features of the same audio sample in a symmetrical manner. On the data enhancement channel of the original audio, in order to reduce the workload of the computer, the original scheme uses random audio scaling instead of audio time stretch, but this method can cause partial audio distortion. In order to reduce the loss of audio acoustic characteristics caused by data enhancement, two methods of audio time stretch and pitch shift processing are adopted to apply to the data enhancement channel of the original audio. This scheme can generate new audio variants without abandoning the audio characteristics, and does not significantly increase the computational burden. On the data enhancement channel of the spectrogram, an adaptive and novel enhancement strategy is proposed for the L-mHP feature, that is, separate data enhancement for each channel. In this way, by randomly covering each channel of the L-mHP features, the model is intended to focus more on learning the time–frequency mode in the spectrogram during the training process, rather than relying on some special image features for classification learning like the image classification method.

The overall working flow chart of the improved dual-channel data enhancement scheme based on a symmetric structure is shown in Figure 3. From this figure, the symmetrical structure uses the feature extraction module as the intermediate module. The first half is the data enhancement channel applied to the original audio, and the latter half corresponds to the data enhancement part applied to the spectrogram features. Specifically, the original audio is first sent into the original audio data enhancement channel to generate audio variants, then the Log-Mel spectrogram features of audio variants are extracted. Next, the harmonic spectrogram and percussive spectrogram are obtained by HPSS on the Log-Mel spectrogram features. The Log-Mel spectrogram, harmonic spectrogram, and percussive spectrogram are sent into the spectrogram feature data enhancement channel for covering processing. Finally, after covering, the three spectrograms are obtained to form the L-mHP features. Thus, the audio dual-channel data process is completed.

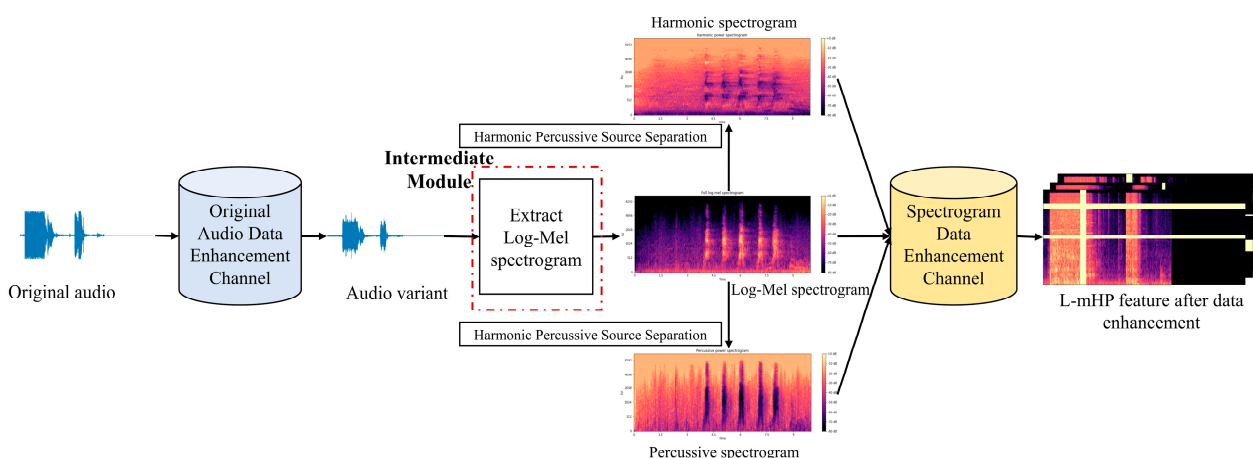


Figure 3. The overall working flow chart of the improved dual-channel data enhancement scheme based on a symmetric structure.

There are several processes in the original audio data enhancement channel, they are Remove Audio Silence, Audio Time Stretch, Audio Pitch Shift [16], and Random Padding. These methods play a very important role in the overall process, so it is necessary to illustrate these parts clearly and precisely.

- A. Remove Audio Silence: There will be a probability that some silent fragments are combined with the endpoints when finishing the recording and interception of the sound signals. These parts may have little key or essential information. After removal, the non-silent part can be adjusted freely along the time axis, resulting in richer variants.
- B. Audio Time Stretch: This method will slightly stretch the audio time while keeping the sound pitch unchanged, generating different samples under the same label. The maximum stretch ratio is 1.5.
- C. Audio Pitch Shift: This method will slightly shift the audio pitch while keeping the sound time rate unchanged. The maximum shift ratio is 1.5. Combining these two approaches results in more flexible variants without losing audio characteristics.
- D. Random Padding: This method refers to the output of the data-enhanced audio at a fixed length. For the ESC-50/10 datasets, the length of the audio sequence is fixed at 220,500 (about 5 s). For the UrbanSound8K dataset, the length of the audio sequence is fixed at 141,120 (about 3.2 s). The specific enhancement process of the channel is shown in Figure 4.

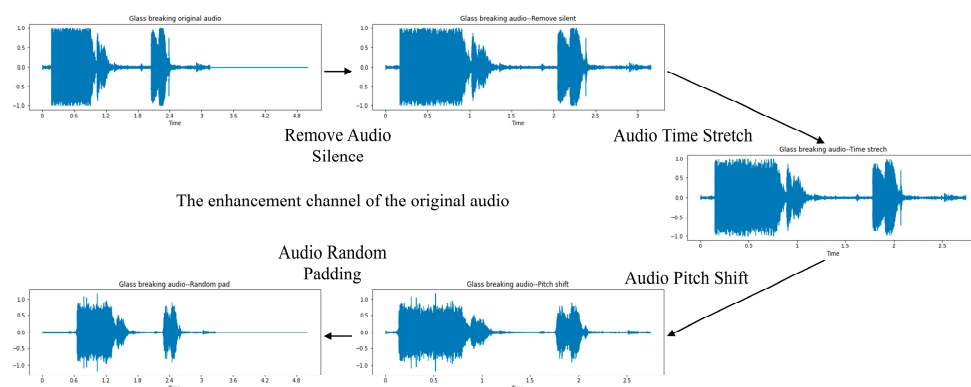


Figure 4. The original audio data enhancement channel process.

The spectrogram data enhancement channel refers to randomly covering the three spectrograms in the L–mHP feature. Specifically, one or more covering bands are set along the time dimension and the frequency dimension, respectively. The bandwidth is an integer randomly selected between 0 and F , $[0, F]$, and the value of the frequency covered by the covering band is set to 0. Then the generated variant forms a three-channel feature for model learning and training. However, it may lead to a lack of more feature information in the spectrogram if the bands are too wide or it contains too many covering bands. It is of no help to the model training and leads to the unexpected learning period of the sound signals. Therefore, it was decided to set one covering band along the frequency dimension and two covering bands along the time dimension, and set the maximum bandwidth to $F = 16$. The channel is shown in Figure 5. During each training run, the variants created for the same samples will have subtle differences due to the randomness of the data enhancement process. These circumstances help the model learn more useful features for ESC tasks. As for the spectrogram data enhancement channel, it can be considered as a random coverage of the features in each channel. In this way, the model is supposed to pay more attention to the time–frequency pattern in the spectrogram environment during the training process. The generalization performance of the model is improved at the same time.

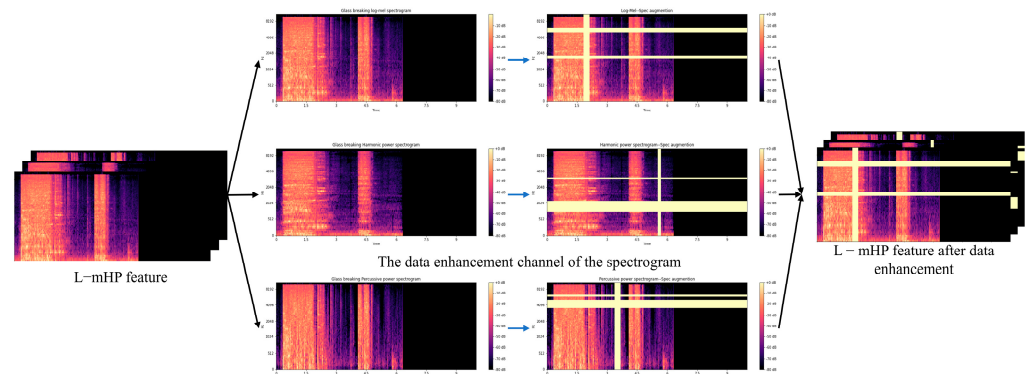


Figure 5. The spectrogram data enhancement channel process.

Subsequent experiments on the baseline model also prove that the improved dual-channel data enhancement scheme based on a symmetric structure makes the audio variants more diverse without occupying more computing resources. It also makes the model focus on learning important sound characteristics in the time domain and frequency domain during the training process. These updating strategies are both helpful to improve the generalization performance and classification accuracy of the model.

The above schemes are implemented by Librosa.

2.3. SE-ResNet50 Classification Model Based on SE Attention Block

2.3.1. SE Attention Block

The SE block is a channel attention mechanism proposed by [25]. Its core idea is to focus on feature channels that are useful for classification tasks through autonomous learning during training, while ignoring useless channel information. The main framework of the SE attention block is shown in Figure 6.

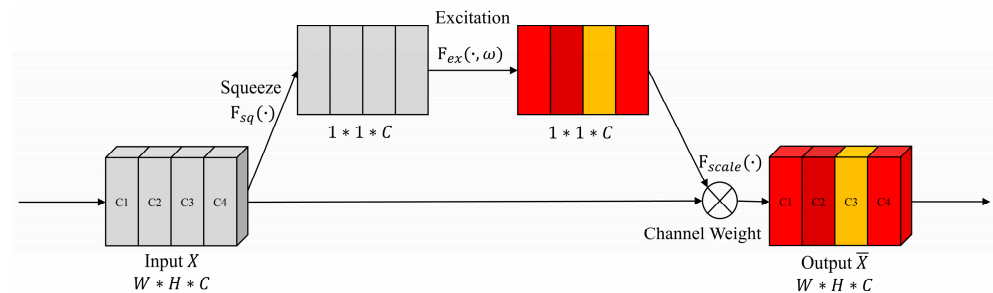


Figure 6. The main framework of SE attention block.

The main components of the SE attention block are the two operations of SE. Firstly, the input feature X is compressed by $F_{sq}(\cdot)$:

$$z_C = F_{sq}(X_c) = \frac{1}{H * W} \sum_{i=1}^H \sum_{j=1}^W X_c(i, j). \quad (7)$$

The squeeze operation is adopted to obtain the global receptive field of the channel feature by using global averaging pooling, and then it performs an excitation operation on z_C by $F_{ex}(\cdot, \omega)$:

$$S = F_{ex}(z, \omega) = \omega_2 \cdot \text{ReLU}(\omega_1 \cdot z). \quad (8)$$

The excitation operation is adopted to use two fully connected layers to predict the importance of each channel feature, where ω_1 and ω_2 represent the network weights of the first fully connected layer and the second fully connected layer, respectively. Then the Sigmoid function is activated to output the adjusted weight of each channel. Finally, the

weighting of the feature channel is completed by an $F_{scale}(\cdot)$ operation, and the output feature $\bar{X}$:

$$\bar{X} = X \cdot F_{scale}(Sigmoid(S)). \quad (9)$$

SE attention blocks in the image field are often applied in structures such as Inception [26], prompting models to learn and model effective content. It is expected to energize the ability of models to extract features.

2.3.2. SE-ResNet50 Classification Model Based on SE Attention Block

As one of the CNN architectures commonly used in the image field, the deep residual network was proposed by He et al. [27] in 2015. As the basic block of the deep residual network, the residuals block is implemented by adding branches to the convolution block, and the branches are used to pass residuals. Its main purpose is to avoid issues, such as gradient disappearance or explosion, as the model depth further deepens. In order to build a deeper network, He et al. proposed the Bottleneck structure on the basis of residual learning. The ResNet-50 network model is composed of two types of Bottleneck structures. The difference is whether there is a Downsample operation in the Bottleneck structure.

The ResNet-50 network model contains a total of 5 Stage blocks. Only Stage 0 contains one convolutional layer, the other 4 Stage blocks are composed of the Bottleneck structures. There are obvious differences among the number of Bottleneck structures inside the Stage blocks. The core idea of the SE-ResNet50 model proposed in this paper is to embed the SE attention block on the basis of the original Bottleneck structure to obtain the SE-Bottleneck structure, and then the SE-ResNet50 model is constructed by using two types of SE-Bottleneck structures according to the structural distribution of the standard ResNet-50 model. This scheme is implemented here to make each SE-Bottleneck structure of the audio classification model pay more attention to the channel information, which is often considered to be helpful for the classification task during the model training process. This part is expected to strengthen the correlation between deep features and shallow features. In this way, the SE-ResNet50 model is composed of these SE-Bottleneck structures that are embedded with an SE attention mechanism, which makes the model learn how to use the important channel features in the acoustic features to complete the classification task, and improves the ability of the whole model to extract key features. The main structure of the model is shown in Figure 7. The SE-Bottleneck1 structures have Downsample operation, and SE-Bottleneck2 is without Downsample operation.

The data-enhanced L-mHP features are sent to the SE-ResNet50 model as input features for training or verification. It first passes through the SE-ResNet50 main module. The feature extraction module can be considered the core part, which is applied to extract high-dimensional information precisely. It will focus more on the acoustic features in a real-world situation and bring necessary benefits to the classification module. The high-dimensional acoustic information extracted by the feature extraction module is then fed into the average pooling layer for feature dimension reduction. Lastly, the acoustic features, after pooling processing, are sent into a fully connected layer for linear transformation, and the prediction sequence of the sound label is attainable. When an input feature X passes through an SE-Bottleneck structure without Downsample operation, the output Y_X is:

$$\begin{aligned} y_1 &= f(B(Con_1(X))), \\ y_2 &= f(B(Con_2(y_1))), \\ y_3 &= B(Con_3(y_2)), \\ &\text{and} \\ Y_X &= f(\sigma\{F_{c2}[f[F_{c1}[GAP(y_3)]]\} + X), \end{aligned} \quad (10)$$

where y_1 , y_2 , and y_3 correspond to the output of the three convolutional layers in the SE-Bottleneck, respectively, $B(\cdot)$ is the batch normalization operation, $f(\cdot)$ is the ReLU

activation function, $GAP(\cdot)$ is the global average pooling, $F_{c1}(\cdot)$ and $F_{c2}(\cdot)$ are two fully connected calculations in the SE attention block, and $\sigma(\cdot)$ is the Sigmoid activation function.

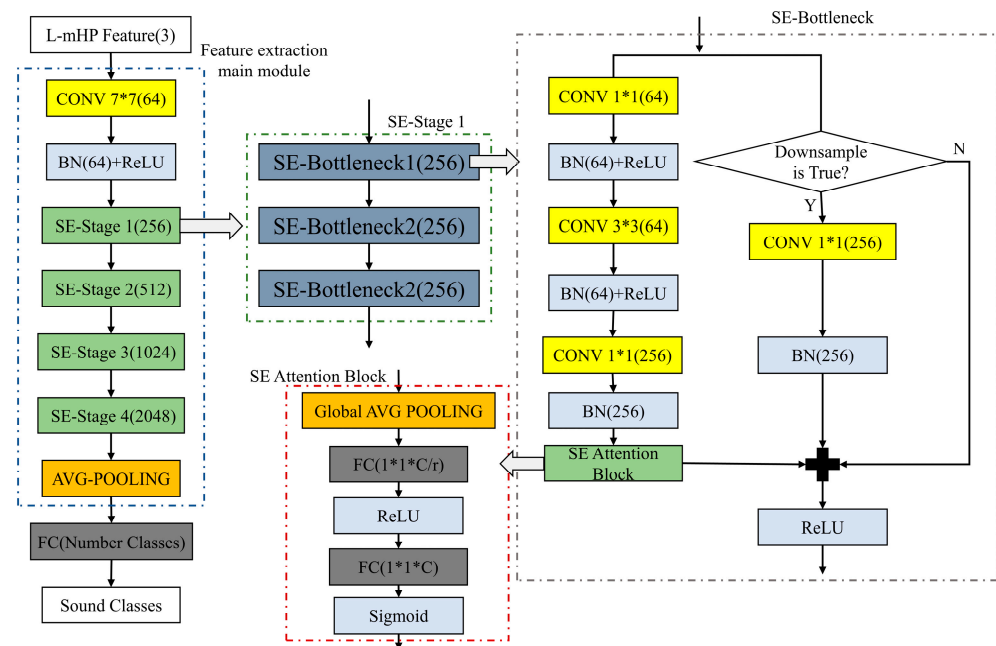


Figure 7. The main structure of SE-ResNet50 model.

3. Experiment Setup

3.1. Public Environment Sound Datasets

Three public datasets, ESC-50, ESC-10 [28], and UrbanSound8K [29], were selected for training and evaluating classification models in this paper. These three datasets are also the datasets commonly used for ESC tasks.

The ESC-50 dataset consists of 2000 sound samples with a fixed length of 5 s, which includes five sound categories, including animal, natural soundscape and water sound, non-verbal human voice, indoor/home sound, and external/urban sound. On this basis, they can be more carefully divided into 50 sound sub-categories, including dog barking, knocking at the door, helicopter sound, wind sound, and so on. The data set is officially divided into five folders, each of which contains an equal number of sounds and a balanced distribution of types.

As a subset of ESC-50, the ESC-10 dataset has 400 sound samples, the length of each sound sample is also 5 s. The ESC-10 includes 10 types of sounds such as barking, rain, waves, baby crying, sneezing, helicopters, and so on. Like ESC-50, ESC-10 is also officially divided into five folders.

UrbanSound8K, later referred to as US8K, has 8732 sound samples from 10 different categories, including air conditioner, car horn, dog bark, children playing, siren, engine idling, gun, street music, and so on. The length of the sound segments is different, and the maximum length is 4 s. In the US8K sound clips, the sampling rate is different, the number of sound channels is not uniform. In the process of data preprocessing, we calculate the average value of the stereo sound channel.

When these three datasets are used in the experiment, all audios are resampled at a sampling rate of 44.1 kHz. The length of the audio in the US8K dataset is uniformly intercepted as 3.2 s during data preprocessing.

3.2. Feature Extraction

When extracting Log-Mel acoustic features, the frame length is set to 1024, the length of overlapping area is 512, and the number of Mel bands is set to 128. During our experiments, all the audio samples are resampled, and the sampling rate is set to 44.1 kHz. On this

basis, the Log-Mel acoustic features are subjected to the HPSS operation to form the L-mHP three-channel features. For the ESC-50/10 datasets, the sample length is fixed for 5 s, and the extracted L-mHP feature dimension is [3, 128, 431]. For the US8K dataset, the sample length is intercepted as 3.2 s in the data preprocessing stage, and the extracted L-mHP feature dimension is [3, 128, 275]. The official segmentation strategy is adopted by US8K.

3.3. Transfer Learning

The acoustic feature adopted in this paper is a three-channel feature L-mHP, which is mainly composed of a Log-Mel spectrogram, a harmonic spectrogram, and a percussive spectrogram. During its core process, the harmonic spectrogram and the percussive spectrogram are both obtained by HPSS of the Log-Mel spectrogram. This method will bring a three-dimensional image feature mapping process for each audio wav file in the dataset. By introducing a transfer learning method and applying it to the feature extraction of the model, state-of-the-art performance is obtained in the ESC task after training and fine-tuning with sound datasets [30]. The idea of transfer learning is to use the knowledge in previously learned models and apply them to later models. The SE-ResNet50 model proposed in this paper is based on the ResNet-50 model pre-trained on the ImageNet dataset and embedded with the SE attention block.

3.4. Baseline Model

The pre-trained ResNet-50 model is selected as the baseline model in this paper, whose input feature turns out to be the feature of copying the Log-Mel spectrogram into three channels. In the training process, a cross entropy loss function is used as the loss function, the batch-size is set to 64, the highest iteration Epoch is 400, and Adam is selected as the parameter optimizer [8–18].

The experimental environment is a Core i7-8550U with an RTX 2080Ti (11 GB) graphics card. All our experiments are performed in Pytorch 1.10.0 using Python 3.8.10.

4. Results and Analysis

4.1. The Transfer Learning Method on Model Performance

Since the baseline model and its improved SE-ResNet50 model in this paper involve transfer learning, in order to verify the impact of transfer learning on classification accuracy, the following ablation experiments are performed on the baseline model. The input feature is to transform the Log-Mel spectrogram into a three-channel feature by replication. The dataset used is ESC-50. The experimental results are shown in Figure 8, and the result visualization tool is provided by TensorBoard 2.7.0.

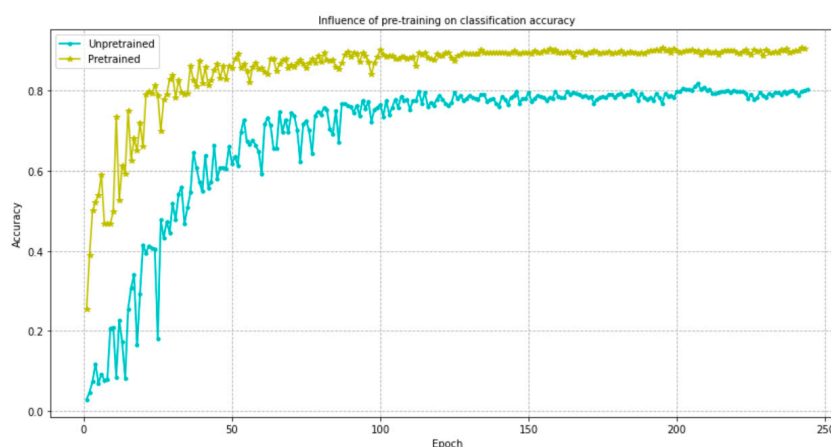


Figure 8. The effect of introducing transfer learning on model performance.

It can be seen from Figure 8 that the ResNet-50 model pre-trained by the ImageNet dataset is far superior to the unpre-trained ResNet-50 model in terms of model convergence

speed and experimental accuracy. Therefore, by introducing a transfer learning method in the training process, the ability of the model to extract deep features can be greatly improved and the convergence speed of the model can be accelerated.

4.2. The Acoustic Features Influence on Experimental Performance

To explore the influence of different input features on the accuracy of the experiment and verify the superiority of the L-mHP feature proposed in this paper, the ablation experiment was carried out on the baseline model, and the following are the feature schemes set up for the control experiment:

- A. Log-Mel spectrogram copy three channels
- B. Log-Mel spectrogram with two zero-value channels
- C. Log-Mel spectrogram with harmonic spectrogram and zero value channel
- D. Log-Mel spectrogram with percussive spectrogram and zero value channel

The experimental results are shown in Figure 9.

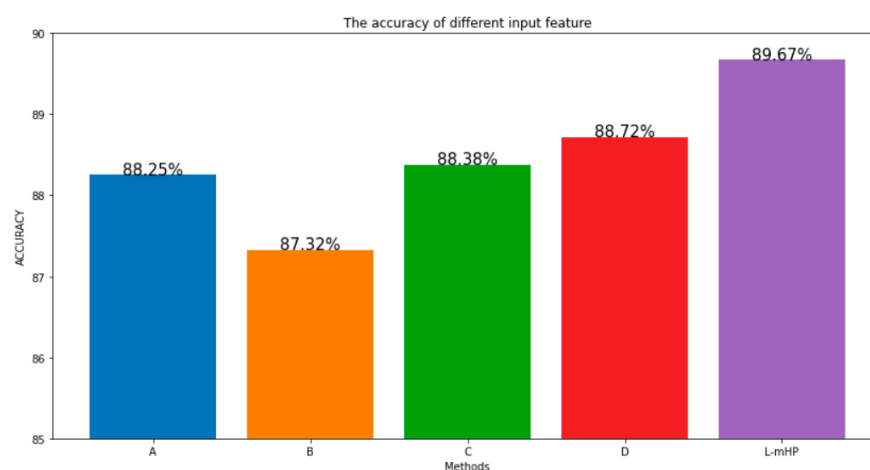


Figure 9. The classification effect of different acoustic feature combinations. Each color corresponds to a feature scheme, respectively.

It can be seen from the experimental results that the classification accuracy of the L-mHP feature proposed in this paper is better than the other three methods on the baseline model, which proves that the feature is more effective in characterizing environmental sounds.

4.3. Affect of the Improved Dual-Channel Data Enhancement Scheme on the Experimental Performance

To investigate whether the improved dual-channel data enhancement scheme based on a symmetric structure is more helpful to the classification task or not, a simulation experiment is carried out on the baseline model. The dataset is ESC-50, and the input feature is the L-mHP feature. The experimental results are shown in Table 1.

By analyzing the experimental results, it can be seen that the improved dual-channel data enhancement scheme and a symmetrical structure in this paper enhance the original audio and spectrogram feature processing method, which makes the audio-generated variants more flexible and diversified. To a certain extent, it can benefit the model from learning more useful feature information for audio classification tasks and improve the performance of the model. At the same time, the additional time of data processing is completely acceptable.

4.4. Influence of the SE Attention Mechanism on Experimental Performance

In order to explore the influence of introducing the SE mechanism on the performance of the classification model, and to verify the effectiveness of using the L-mHP feature to

represent environmental sound and using an improved dual-channel data enhancement scheme based on a symmetric structure for model training, which are proposed in this paper to improve experimental performance, relevant ablation experiments are carried out using the ESC-50 dataset. The specific experimental results are shown in Table 2.

Table 2. Ablation experiments on SE attention mechanism.

Classification Model (Input Feature)	Data Enhancement Scheme	Accuracy (%)
ResNet-50 (three-channel Log-Mel)	Just Spectrogram	88.25
ResNet-50 (L-mHP)	Just Spectrogram	89.67
SE-ResNet50 (L-mHP)	Just Original Audio	92.2
SE-ResNet50 (L-mHP)	Just Spectrogram	93.05
SE-ResNet50 (L-mHP)	All (dual-channel)	94.92

From Table 2, it can be seen that the baseline model combined with the three-channel Log-Mel spectrogram, and the data enhancement scheme are only applied to the spectrogram. This scheme achieves 88.25% classification accuracy. After replacing the L-mHP feature as the input feature, the classification accuracy of the model is improved by 1.42%. When the SE-ResNet50 model is used as the classification model, the L-mHP feature is used as the input feature, and the data enhancement scheme is only applied to the original audio, the classification accuracy reaches 92.2%, which is 3.95% higher than the baseline model. Based on the above scheme, when only the data enhancement scheme is adjusted and applied to the spectrogram, the experimental accuracy is 93.05%, which is 4.8% higher than the baseline model and 0.85% higher than the scheme that just uses audio enhancement. This shows that the positive feedback obtained by the data enhancement scheme for the spectrogram is slightly better than the data enhancement scheme only for the original audio. Finally, adopting the audio classification framework proposed in this paper, which is the SE-ResNet50 model and the L-mHP feature, combined with the improved dual-channel data enhancement scheme based on a symmetric structure, a classification accuracy of 94.92% can be obtained, which is 6.67% higher than the baseline model and the highest accuracy in all ablation experiments. Through the above experiments, it is proved that both the L-mHP feature, the improved dual-channel data enhancement scheme, and the SE attention block are helpful to improve the classification accuracy.

4.5. Experimental Performance Comparison of Different Audio Classification Frameworks

The audio classification framework proposed in this paper is composed of L-mHP features as input features, and the SE-ResNet50 network model as a classification model, with an improved dual-channel data enhancement scheme based on a symmetric structure. In order to verify the classification performance of the framework, experiments are carried out on the public data sets ESC-50, ESC-10, and US8K. They were adopted to be compared with some current mainstream models. The experimental results are shown in Table 3.

Table 3. Comparative experimental results of different audio classification models (%).

Classification Model	Architecture (Input Feature)	ESC-10	ESC-50	US8K
Piczak-CNN [14]	2 CONV + 2 FC (Mel spectrogram)	90.20	64.50	73.70
EnvNet-v2 [31]	10 CONV + 3 FC + 5 POOL (wave)	91.30	84.70	78.30
Baseline	Pre-trained ResNet-50 (Log-Mel spectrogram)	92.75	88.25	83.37
ESResNet-A [9]	ResNet-50 + Self attention (Log-STFT spectrogram)	97.00	91.50	85.42
DenseNet [18]	DenseNet-201 (Log-Mel spectrogram)	-	91.25	85.31
SoundCLR [10]	ResNet-50 + Hybrid Loss (Log-Mel spectrogram)	97.85	92.25	86.49
LSCNet [32]	LSCNet (Mel + HPSS)	-	88.00	90.10
Our	ResNet-50 + SE Attention (L-mHP feature)	99.67	94.92	91.75

The baseline model uses the ResNet-50 model. After pre-training with the image datasets, this model can obtain relatively good accuracy in the audio classification task by fine-tuning parameters. The ResNet-50 model with three-channel Log-Mel spectrogram as the input feature, on the public datasets ESC-50, ESC-10, and US8K, reached 88.25%, 93.75%, and 83.37%, respectively. It can be vividly seen that its performance is superior to some traditional audio classification models [14,31]. The classification framework proposed in this paper has achieved 94.92%, 99.67%, and 90.75% experimental accuracy on ESC-50, ESC-10, and US8K, respectively. Based on the baseline model, it has increased by 6.67%, 6.92%, and 8.38%, respectively, which is better than some current mainstream audio classification models. Figure 10 shows the confusion matrix of the audio classification framework applied to the ESC-50 dataset. The SE attention block is introduced into the ResNet-50 model, which boosts the model to focus on the feature channels that are useful for the classification task and ignore the useless feature information in the training process. This is then combined with the L-mHP feature and the improved dual-channel data enhancement scheme. This state-of-the-art method has a tremendous classification effect. The classification model used in reference [9] also introduces a self-attention mechanism based on the ResNet-50 model. Table 4 briefly evaluates aspects of the two kinds of attention: experimental accuracy, parameters, and FLOPs.

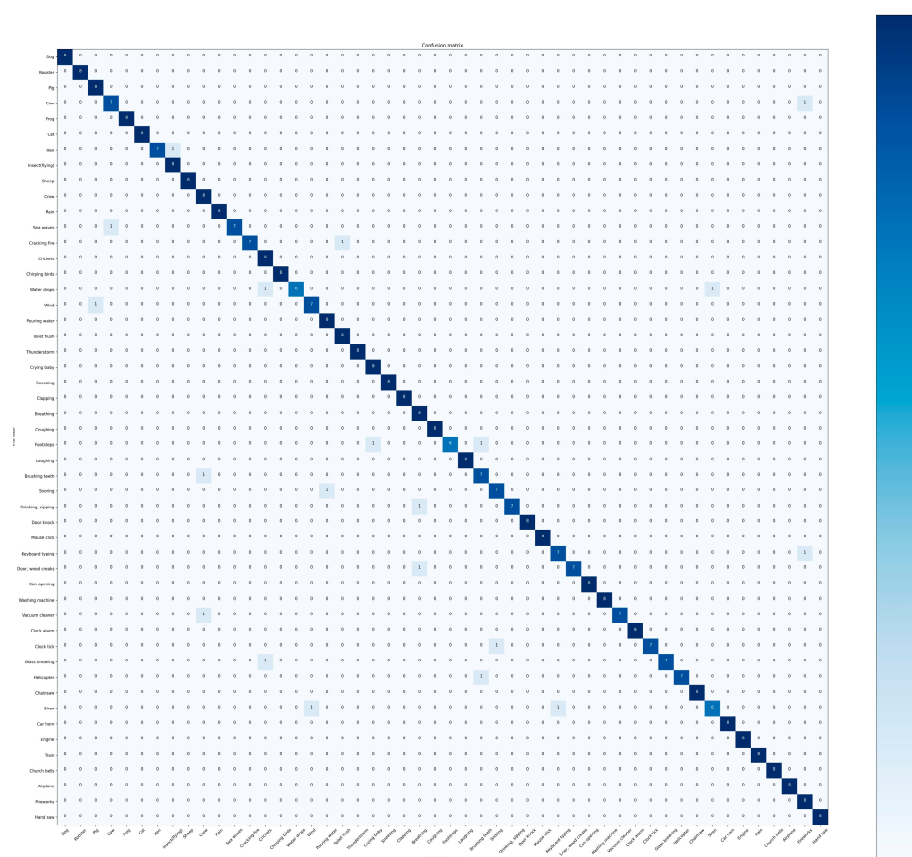


Figure 10. The confusion matrix obtained by the proposed framework on ESC-50.

Table 4. Comparison of two models with attention mechanism.

Classification Model	Accuracy (%)	Parameters (M)	FLOPs (G)
ResNet-50 (Baseline)	88.25	23.51	4.57
ESResNet-A [9]	91.50	30.53	4.98
Our	94.92	26.02	4.58

The ESResNet-A [9] model introduces a self-attention mechanism into the ResNet-50 model, and its core idea is to use deep separable convolution to extract the residuals of features. The purpose of the SE attention block is to make the model focus more on essential information channels through compression and excitation operations. It is then intended to be combined with the L-mHP features and the improved dual-channel data enhancement scheme, so that the classification framework is superior to the ESResNet-A model in terms of parameter quantity, computational complexity, and classification accuracy.

4.6. Self-Made Sound Datasets

Although the audio classification framework proposed in this paper is trained and evaluated on three public sound datasets, the actual effect on a real-world situation should be considered. The sound signals' set may contain more noise, and the audio quality be worse, in a real-world situation. Therefore, to evaluate the performance of the audio classification framework in the actual environment, it was decided to simulate the sound samples in an actual environment through a self-made dataset. A total of 200 audios were extracted from the ESC-50 dataset, including dog barking, car horn, glass breaking, siren, and engine sound. Noise was added randomly at four signal-to-noise ratios of -5 , 0 , 5 , and 10 in each of the five classes of sounds. The noise included rain and wind, and each noise was selected with a probability of 50% , and the noise was also from the public sound dataset. In this way, each audio generated three variants, corresponding to four signal-to-noise ratios. The self-made dataset was divided into four folders according to the signal-to-noise ratio, and each folder contained 200 audios. The self-made dataset was sent to the baseline model, ESResNet-A, and the audio classification framework proposed in this paper, for evaluation and analysis. The experimental results are shown in Figure 11.

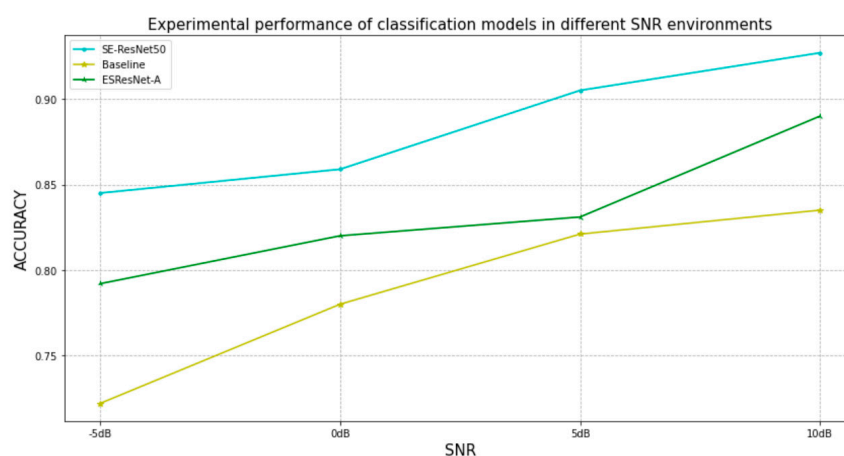


Figure 11. Classification results of datasets with different signal-to-noise ratios on three audio classification models.

It can be seen from Figure 11 that the audio classification framework proposed in this paper, that is, the classification model SE-ResNet50, combined with L-mHP acoustic features, and the improved dual-channel data enhancement scheme, gave the most accurate results. A classification accuracy of 0.8450 , 0.8589 , 0.905 , and 0.927 can still be achieved on the self-datasets whose signal-to-noise ratios are -5 dB, 0 dB, 5 dB, and 10 dB, respectively, and the model performance is better than the other two models. It is proved that the SE-ResNet50 network model, L-mHP acoustic features, and improved dual-channel data enhancement scheme based on a symmetric structure mentioned in this paper are of great significance to improve the classification performance and enhance the generalization ability of the model. It also proves the application feasibility and value of the classification framework in real-life scenarios.

5. Conclusions

A novel audio classification framework is proposed based on a L-mHP feature and SE-ResNet50 model in this paper. Meanwhile, a dual-channel data enhancement scheme based on a symmetric structure is improved for model training. The experimental results on public datasets show that the proposed audio classification framework has excellent classification performance. At the same time, the evaluation experiments on a self-made dataset show that the proposed method is robust and is highly feasible for application in the actual environment. The ablation experiments show that the L-mHP feature proposed here can adequately characterize environmental sound. The SE-ResNet50 model can focus on important feature channels independently in the training process by embedding an SE attention block, and the feature extraction ability is greatly improved. The proposed methods combined with an improved dual-channel data enhancement scheme gives great performance in ESC tasks.

However, the audio classification framework has limited ability to deal with multiple overlapping sounds. Alternatively, the classification accuracy may be affected significantly by a more severe sound environment, such as heavy rain or strong winds.

In future work, we will focus on the classification of overlapping sounds and low signal to noise sound signals. We will next investigate how to apply the noise reduction algorithm to the classification task and how to combine the sound separation algorithm when classifying overlapping sound events etc.

Author Contributions: Conceptualization, M.H. and M.W.; software, M.H. and X.L.; experiment analysis, M.H. and R.K.; writing—review and editing, M.H., M.W. and H.Q.; funding acquisition, M.W. and H.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is supported by the National Natural Science Foundation of China under Grant 62071135 and 61961010. It is also supported by Innovation Project of GUET Graduate Education No. 2023YCXB05.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the grants for this manuscript are still in the research phase, and some research data or key codes are currently limited to disclosure within the project team. However, the self-made datasets of the environment sound involved in this manuscript can be provided. If necessary, you can contact Mengxiang Huang via email (1020210969@glut.edu.cn) to obtain the Baidu Netdisk (Baidu Cloud) URL link and then download the files you need. Once the link is gone, you can contact the authors of this article to obtain the latest link.

Acknowledgments: We are very grateful to volunteers from GLUT for their assistance in the experimental part of this manuscript.

Conflicts of Interest: The authors declare that there is no conflict in interest or personal relationship.

References

1. Jin, Q.; Liang, J. Video Description Generation using Audio and Visual Cues. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR'16). Association for Computing Machinery, New York, NY, USA, 6–9 June 2016; pp. 239–242.
2. Jiang, D.-N.; Lu, L.; Zhang, H.-J.; Tao, J.-H.; Cai, L.-H. Music type classification by spectral contrast feature. In Proceedings of the IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, 26–29 August 2002; Volume 1, pp. 113–116.
3. Cotton, C.V.; Ellis, D.P.W. Spectral vs. spectro-temporal features for acoustic event detection. In Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 16–19 October 2011; pp. 69–72.
4. Li, R.; Yin, B.; Cui, Y.; Du, Z.; Li, K. Research on Environmental Sound Classification Algorithm Based on Multi-feature Fusion. In Proceedings of the 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 11–13 December 2020; pp. 522–526.
5. Wu, J.; Gao, Z. D-S sound classification based on double two-stream convolution and multi-feature fusion. *Appl. Res. Comput.* **2022**, *39*, 7.

6. Zhang, K.; Su, Y.; Wang, J.; Wang, S.; Zhang, Y. Environment Sound Classification System Based on Hybrid Feature and Convolutional Neural Network. *J. Northwestern Polytech. Univ.* **2020**, *38*, 162–169. [\[CrossRef\]](#)
7. Shao, Y.; Wang, D. Robust speaker identification using auditory features and computational auditory scene analysis. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 1589–1592.
8. Xu, W.; Zhang, X.; Yao, L.; Xue, W.; Wei, B. A multi-view CNN-based acoustic classification system for automatic animal species identification. *Ad. Hoc. Netw.* **2020**, *102*, 102115. [\[CrossRef\]](#)
9. Guzhov, A.; Raue, F.; Hees, J.; Dengel, A. ESResNet: Environmental Sound Classification Based on Visual Domain Models. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4933–4940.
10. Nasiri, A.; Hu, J. SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification. *arXiv* **2021**, arXiv:2103.01929.
11. Zhang, Z.; Xu, S.; Cao, S.; Zhang, S. Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. In *Pattern Recognition and Computer Vision. PRCV 2018; Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2018; Volume 11257.
12. Zhang, Z.; Xu, S.; Zhang, S.; Qiao, T.; Cao, S. Learning Attentive Representations for Environmental Sound Classification. *IEEE Access* **2019**, *7*, 130327–130339. [\[CrossRef\]](#)
13. Su, Y.; Zhang, K.; Wang, J.; Zhou, D.; Madani, K. Performance analysis of multiple aggregated acoustic features for environment sound classification. *Appl. Acoust.* **2020**, *158*, 107050. [\[CrossRef\]](#)
14. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
15. Bahmei, B.; Birmingham, E.; Arzanpour, S. CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2022**, *29*, 682–686. [\[CrossRef\]](#)
16. Peng, N.; Chen, A.; Zhou, G.; Chen, W.; Zhang, W.; Liu, J.; Ding, F. Environment Sound Classification Based on Visual Multi-Feature Fusion and GRU-AWS. *IEEE Access* **2020**, *8*, 191100–191114. [\[CrossRef\]](#)
17. Palanisamy, K.; Singhanian, D.; Yao, A. Rethinking CNN Models for Audio Classification. *arXiv* **2020**, arXiv:2007.11154.
18. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [\[CrossRef\]](#)
19. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, November 2015; pp. 1–16. Available online: <https://arxiv.org/abs/1511.06434v2> (accessed on 2 May 2016).
20. Zhao, W.; Yin, B. Environmental sound classification based on adding noise. In Proceedings of the 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 17–19 December 2021; pp. 887–892. [\[CrossRef\]](#)
21. Allen, J. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1977**, *25*, 235–238. [\[CrossRef\]](#)
22. Jabnouni, H.; Arfaoui, I.; Cherni, M.A.; Bouchouicha, M.; Sayadi, M. ResNet-50 based fire and smoke images classification. In Proceedings of the 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sfax, Tunisia, 24–27 May 2022; pp. 1–6. [\[CrossRef\]](#)
23. Xu, J.-X.; Lin, T.-C.; Yu, T.-C.; Tai, T.-C.; Chang, P.-C. Acoustic Scene Classification Using Reduced MobileNet Architecture. In Proceedings of the 2018 IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 10–12 December 2018; pp. 267–270.
24. Akaishi, N.; Yatabe, K.; Oikawa, Y. Harmonic and Percussive Sound Separation Based on Mixed Partial Derivative of Phase Spectrogram. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 301–305.
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
26. Wang, H.; Huo, W.; Lu, Y.; Pei, J.; Huang, Y. Modulation Recognition of Overlapping Radar Signals under Low SNR Based on Se-Incepatnet. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2734–2737. [\[CrossRef\]](#)
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM international conference on Multimedia (MM'15), Brisbane, Australia, 26 October 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1015–1018.
29. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia (MM'14), Orlando, FL, USA, 3–7 November 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 1041–1044.

30. Bi, S.; Xu, L.; Zhao, S.; Wang, J. Acoustic Scene Classification for Bone-Conducted Sound Using Transfer Learning and Feature Fusion. In Proceedings of the 2022 5th International Conference on Information Communication and Signal Processing (ICICSP), Shenzhen, China, 26–28 November 2022; pp. 519–522.
31. Tokozume, Y.; Ushiku, Y.; Harada, T. Learning from Between-class Examples for Deep Sound Recognition. *arXiv* **2017**, arXiv:1711.10282.
32. Fei, H.; Wu, W.; Li, P.; Cao, Y. Acoustic scene classification method based on Mel-spectrogram separation and LSCNet. *J. Harbin Inst. Technol.* **2022**, *54*, 124.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.