A two-stream convolution architecture for environment sound classification based on audio feature disentanglement

ZhengHao Chang 1347712590@QQ.COM
Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion (Wuhan Textile University), Wuhan 430200, China.

RuHan He* HERUHAN@WTU.EDU.CN
Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion (Wuhan Textile University), Wuhan 430200, China.

YongSheng Yu YONGSHENGYU@WHUT.EDU.CN
State Key Laboratory of Silicate Materials for Architectures, Wuhan University of Technology, Wuhan, 430070, China.

ZiLi Zhang ZLZHANG@WTU.EDU.CN
School of Computer Science and Artificial IntelligenceWuhan Textile University

GeLi Bai 67799484@QQ.COM
College of Computer and Information Engineering, Inner Mongolia Agricultural University

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Abstract: ESC (Environmental Sound Classification) is an active area of research in the field of audio classification that has made significant progress in recent years. The current mainstream ESC methods are based on increasing the dimension of the extracted audio features and therefore draw on the two-dimensional convolution methods used in image processing. However, two-dimensional convolution is expensive to train and the complexity of the corresponding model is usually very high. In response to these issues, we propose a novel two-stream neural network model by the idea of disentanglement, which uses one-dimensional convolution for feature extraction to disentangle the audio features into the time and frequency domains separately. Our approach balances computational pressure with classification accuracy well. The accuracy of our approach on the Urbansound 8k and Esc-10 datasets was 98.51% and 97.50%, respectively, which exceeds that of most models. Meanwhile, the model complexity is also lower.

Keywords: environmental sound classification, audio feature, feature disentanglement, two-stream neural network

1. Introduction

In the real world, we are surrounded by sounds from the environment and are constantly on receiving audio signals. Sound signals contain a wealth of information that would be of great help to humanity if machines were made to understand them properly. Thanks to deep learning, speech recognition technologies have become increasingly mature, especially automatic speech recognition(ASR) Khamparia et al. (2019) and music information recognition(MIR) **?**. Environmental sound classification(ESC) is more challenging than ASR and

MIR Liu et al. (2021). Furthermore, the semantic annotation of soundscapes or acoustic scenes is an open task, as no complete classification can cover all possible environmental categories Su et al. (2019). ESC has been considered as a supervised recognition task in a closed domain of visible categories. However, such samples are often skewed by environmental factors Barchiesi et al. (2015). The classification task becomes complicated when acoustic events overlap. Thus, the development of an efficient ESC method is more challenging.

The ESC is typically composed of two basic components: feature extraction and classifier design. The former part of ESC is mainly manually extracted features such as mel frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), short-term energy and over-zero rate Davis and Mermelstein (1980).

The later part of ESC is various classifiers, which can be divided into three main categories according to the input feature. In the first category method, the network is trained using the raw audio signal. Tokozume et al. (2017) suggested an end-to-end approach to ESC based on a 2D-CNN. Their network structure and learning method can learn long periods of sound without over-fitting.Moreover, they investigated the most suitable number of convolutional layers for raw feature extraction and their optimal filter size. Dai et al. (2017) proposed a 1D-CNN network that has a wider field of perception and can achieve better results than shallow networks. Abdoli et al. (2019) proposed an end-to-end ESC method based on 1D-CNN. The advantage of this method is that there is no need to extract features manually. However, their 1D-CNN did not account for the temporal structure and frequency characteristics. In the second type of method, the feature spectrum of the original signal is used as input to the model by extracting it. Common features are MFCC and Log-Mel Lim et al. (2018); Abdoli et al. (2019); Dai et al. (2017). Piczak (2015) used Log-Mel features as model input, their proposed ESC system, consisting of two maximum pooling layers, two CNN layers and two fully connected layers and dropout layers, produced classification results that were 5.6% higher than those of machine learning methods. In the third category method, Li et al. (2018) proposed a novel stacked CNN model with multiple convolutional layers of decreasing filter sizes to improve the performance of CNN models with either Log-Mel feature input or raw waveform input. Su et al. (2019) proposed a four-layer convolutional neural network (CNN) that utilized the proposed aggregate features to improve ESC performance. Liu et al. (2021)used a stacked deep network based on aggregation of time-domain features and frequency-domain functions to capture a more comprehensive representation of sound. Their methods have a high accuracy on Urbansound8k. However, multi-stream CNNs are not only complex in structure, but also combine the raw signal with a short time fourier transform, resulting in a large amount of data and high hardware requirement.

Disentangled representation learning used to be discussed at many top conferences and has yielded many achievements, especially in speech recognition and signal processing area Bengio et al. (2013), Bian et al. (2019), Zhang et al. (2019), An et al. (2022)which show the big potential of disentangled representation learning. Therefore, we attempt to introduce disentangled representation learning into environment sound classification.

In this paper, a two-stream convolutional neural network with the idea of disentanglement is proposed. It also takes MFCC as the input, which is commonly used in audio classification. However, different from the other methods, the MFCC features in our method is analyzed in detail with the idea of disentanglement. We investigate that the rows of MFCC

features represent information from the same spatial location on different channels, while the columns represent information from different spatial locations on the same channel. By disentangling the MFCC features in this way, we performed a one-dimensional convolution of the MFCC features. Our method belongs to the third category above. Although our method is a two-stream system, it needes to extract single audio features. Therefore, the cost of feature extraction is similar to that of the second method. Our approach better balances computational pressure with classification accuracy.
In brief, the contribution of this paper is as follows.
• Inspired by the idea of disentanglement, we perform an in-depth analysis of audio features and firstly attempt to treat the time and frequency domains of audio features separately.
• We propose a novel two-stream environment sound classification model and achieve competitive classification accuracy as well as model complexity. Meanwhile, feature extraction brings less computational pressure.
• We investigate and experimentally demonstrate that the validity of the model of our proposed model and the feature of the time domain or frequency domain can serve as the basis for the classification itself, with acceptable classification results.
The remainder of this paper is structured as follows. Section 2 provides an overview of the proposed CNN architecture. Section 3 presents detailed experimental results and a comparison of our classification system with the preceding systems. Section 4 concludes the paper.

2. Method

2.1. Problem Statement

In this research, our task is to construct a deep learning model to achieving automatic environment sound classification. Let $s = (x, y) \in (X \times Y)$ be a training sample pair, where x represents a string of audio signals,and y is the corresponding classification label. Then we create a mapping $\mathbb{R}^x \to \mathbb{R}^f$. The purpose of this mapping is to project an audio signal into a two-dimensional size. So we get $s = (f, y) \in F \times Y$, we try to learn a nonlinear deep learning model(DLM) to classify f as one type of y. The hypothesis of the environment sound classification neural network g_{dlm} is a function which belongs to the hypothesis class G_{dlm}, i.e.$g_{dlm} \in G_{dlm}$. A loss function $l : g_{dlm}(F) \times Y \to R_+$ is proposed, where $l(g_{dlm}(f), y)$ is loss value of g_{dlm} with the sample $s = (f, y)$. Given $D_{\text{train}} = \{s_i = \{f_i, y_i\}\}_{i=1}^{N}$ as a set of training samples, the empirical loss of the classification neural network on D_{train} is defined as:

$$\mathcal{L}_{\text{train}}\ (g_{dlm}) = \frac{1}{N} \sum_{\text{i=1}}^{\text{N}} l\left(g_{dlm}\left(f_i\right), y_i\right) \tag{1}$$

Hence,the classification neural network is trained on the training dataset D_{train} , which aims to optimize the following objective function:

$$\begin{aligned} g_{dlm}^{*}(x) = & \underset{g_{dlm} \in G_{dlm}}{\arg\min}\ \mathcal{L}_{\text{train}}\ (g_{dlm}) + \lambda \mathcal{R}\left(g_{dlm}\right) \\ = & \underset{g_{dlm} \in G_{dlm}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} l\left(g_{dlm}\left(f_i\right), y_i\right) + \lambda \mathcal{R}\left(g_{dlmW}\right) \end{aligned} \tag{2}$$

Where $\mathcal{R}(g_{dlm})$ is a regularization item to modulate the hypothesis function g_{dlm} and λ is a regularization parameter.

2.2. Feature Extraction

Contrary to speech, an environmental sound classification (ESC) is a background sound that is often mixed with various background noises and is therefore harder to recognize. MFCC has been used to solve the automatic speech recognition (ASR) problem, and it does have better performance for artificial speech recognition. Therefore, we decide to introduce MFCC to deal with the ESC problem. For each experiment, we use MFCC feature as input to our models. The hyperparameters in extracting MFCC features are as follows. We set the sample rate of the audio to 44.1 KHZ and framing the input audio with a frame length of 1024 and a hop length of 512 then hanning window for the windowing process. We adopt the number of mel filters M = 128. The frequency response of a triangular filter is defined as:

$$H_m(k) = \begin{cases} 0 & ,k < f(m-1) \\ \frac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & ,(m) \le k \le f(m+1) \\ 0 & ,k \ge f(m+1) \end{cases} \tag{3}$$

Note:$\sum_{m=0}^{M-1} H_m(k) = 1$.

2.3. Proposed network architecture

Inspired by related work in the field of natural image processing, Tolstikhin et al. (2021) achieved a mapping from the vector space of [W,H,C] to the vector space of [S,C]. The rows of the feature Table [S,C] represent information on different channels at the same spatial location, and the columns represent information on the same channel at different spatial locations. In other words, an operation on each row of the table enables the fusion of information in the channel domain, and an operation on each column of the Table enables the fusion of information in the spatial domain. Then they achieve information fusion in both the spatial and channel domains through MLP. We found that current audio features, such as MFCC features, have similar characteristics. We therefore used the above method for classifying the audio functions. Since their MLP method is actually an equivalent form of the convolution operation, we use convolution operations for the extraction of features in the time and frequency domains of audio features respectively. Some studies Kiranyaz et al. (2018); Abdeljaber et al. (2017); Avci et al. (2017) have shown that for some applications, 1D CNNs are advantageous and hence preferable to their 2D counterparts when it comes to 1D signals for the following reasons:

•There is a significant difference in terms of the computational complexities of 1D and 2D convolutions. We compare the mainstream methods of classification using 2D convolution with our proposed method in terms of computational effort. i.e., an input with $W \times H$ dimensions convolve with $K \times K$ kernel and stride with 1 will have a computational effort $E_{2D} = (W - K + 1)(H - K + 1)K^2$ while in the corresponding 1D convolution (with the same dimensions, N ,K) , this is $E_{1D} = (H - K + 1)KW$.

•Using 2D convolution with both frequency and time domain information for audio features may lead to a mishmash of time and frequency domain features into the feature map at Flatten. Our method performs separate feature extraction in the time and frequency domains by means of feature disentangling. We have a more interpretable method.
• The number of parameters in a 1D convolutional neural network is generally much lower than the number of parameters in a 2D convolutional neural network.
So we decided to use 1D convolution for feature extraction. The network structure of the method in this paper is a two-stream 1D convolutional CNN. The first input stream of the model is the MFCC features and the second input stream is the transposed MFCC features, which perform feature extraction for the time-domain and frequency-domain respectively. The time and frequency domain features are fused by global averaging pooling and later fed into 'softmax' classifier for classification. The overall architecture of the model is presented respectively in Fig. 1. As shown in the Fig. 1, we propose an iterable module. The classification accuracy of our model can be slightly increased by stacking more iteration blocks, but more resources will be consumed at the same time. Therefore, the number of the iteration blocks in each stream is set to be 2 in our experiment for trade-off the speed and accuracy. The feature map is sent to this module as input to our proposed convolution block. Our proposed convolution block is shown in the figure Fig 2, Then the output feature map is subjected to the relu activation function operation. The SE Attention Mechanism is then added to the output of the previous step to weight the feature map. In deep learning, large fields of perception generally lead to improved outcomes and extremely costly, especially 2D convolution. When we focus only on a single domain, the cost of getting a broader field of perception is perfectly acceptable. In order to increase the perceptual field of the model, several convolutional kernels of different odd sizes were set up, ranging from 1 to 21. The number of filters multiplied per convolution kernel of different sizes is 16. We added batch normalization (BN) prior to the activation function after the convolution layer. In addition, we did not perform pooling operations, as this would lead to the information loss.

3. Experiment and result

3.1. Experimental setup

3.1.1. Dataset

In this section, we will verify the effectiveness of the method using the ESC-10 and Urbansound8k dataset. Tab 1 provides detailed information about each dataset.
•ESC-10
The dataset contains 400 audios of less than 5 seconds, divided into 10 categories of 40 each, for a total duration of 33 minutes. According to the authors of this dataset, the average human classification rate for this dataset is 95.7%.
•Urbansound8k
The dataset includes 8732 indoor and outdoor audios with a duration of less than 4 s, unevenly divided into 10 categories, for a total duration of 582 minutes.
As the audio durations in the above two datasets are not the same, direct extraction of the original audio features would result in a different dimensionality of the extracted features.

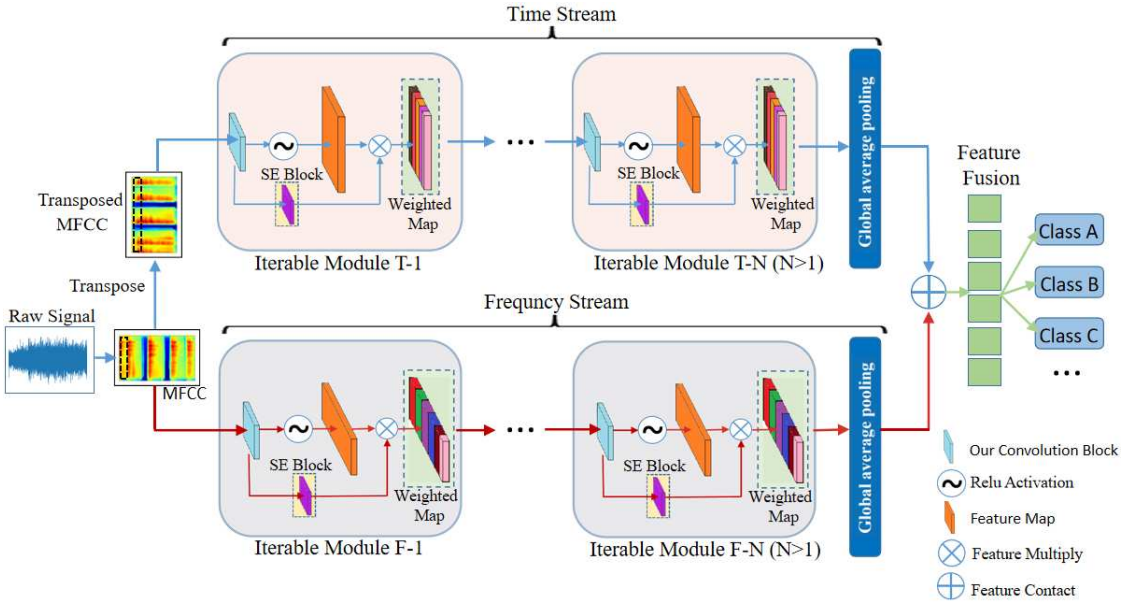


Figure 1: The overall architecture of the proposed model

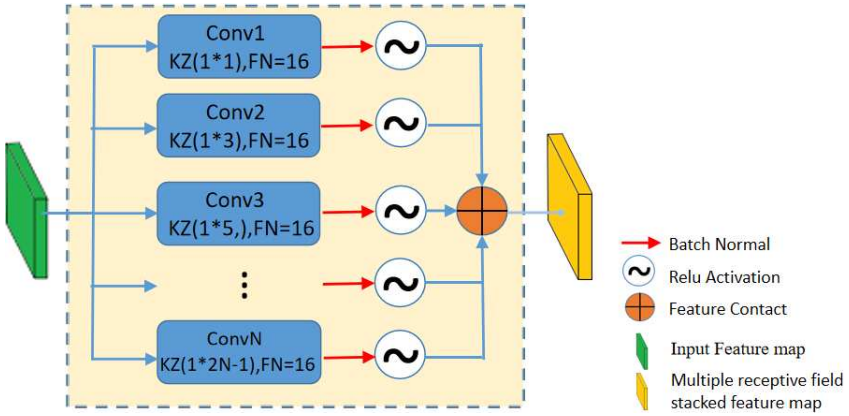


Figure 2: The architecture of our convolution block

In order to remedy this situation, we used the method that proved effective Dong et al. (2020).

Datasets	Classes	Total Duration(mins)	Sample numbers
ESC-10	10	33	400
	dog barking (DB), fire crackling (FC), baby cries (BC), rain (RA), person sneezing (PS), rooster (RO), sea waves (SW), helicopter (HE), chainsaw (CH), and clock tick (CT)		
Urbansound8k	10	582	8732
	air conditioner (AC), dog barking (DB), car horn (CH), children playing(CP), gunshot(GS), engine idling (EI), street music (SM), siren (SI), jackhammer (JA), and drilling (DR)		

Table 1: The description of the used dataset.

3.1.2. Data Augmentation

In deep learning, the number of samples is generally required to be sufficient, the more samples there are, the better the model formed and the more generalizable the model is. The generalizability of the model can be enhanced by increasing the quantity of data for training, and the robustness of the model can be enhanced by adding noisy data. An elegant solution to overcome data scarcity and improve classification performance is data augmentation, where the annotated set of training samples is deformed one or more times to produce dynamic redundant training data Mushtaq and Su (2020); Lu et al. (2017); Zhang et al. (2021). Effective complementary solutions for a short time, audio have also been suggested Dong et al. (2020). Therefore, We choose pitch shift, time stretches and the random padding method. Comprehensive information relating to each of the enhanced datasets is shown in Tab 2.

•Pitch Shifting (PS): The pitch of audio signal clips is shifted by the positive factor of two while keeping the duration unchanged[-2,+2].

•Time Stretching (TS): It is an effective method to increase or decrease audio playback speed. Part of the samples is time-stretched by two factors: [0.91, 1.09].

•Random Padding(RP):Padding shorter lengths of audio at random.

Our training and testing sets wer e randomly split, with training and test sets accounting for 90%t and 10% of the overall data, respectively. To ensure the authenticity of the prediction results, we only augmented the training set.

Datasets	Classes	Total Duration(mins)	Sample numbers after data augmentation
ESC-10	10	122.1	400+1080
Urbansound8k	10	3253.4	8732+23577

Table 2: The Augmented description of the dataset used.

3.1.3. Hardware and software requirements

This experiment was done on Linux version 4.4.0-142-generic operating system, using python 3.6 for the whole experiment. The CPU was 9 Intel(R) Xeon(R) Gold 5218 CPUs @ 2.30GHz. The graphics card used in the system was Tesla V100-PCIE-32GB.
This experiment uses different software and API libraries and packages to train the model from scratch. The operating system used for the experiments was Ubuntu 18.0.4. All experiments were done in python, version 3.6, and the main libraries used were keras2.3.1, tensorflow2.1.0, numpy1.19.5, and librosa0.8.1.
Our optimizer was chosen to be Adam, with an initial learning rate of 0.001, an exponential decay rate of 0.9 for first-order moment estimation and 0.999 for second-order moment estimation, and a total of 500 training epochs. We added dropout and SE Block to the model at the appropriate locations with ratio equals 8.

3.2. Result

This section discusses the performance evaluation of using the model on both the original dataset and the offline enhanced dataset. The evaluation criteria considered are accuracy and the total number of trainable parameters.
We have devised a total of two sets of experiments to evaluate the effectiveness and generalizability of the proposed approach. The first set of experiments uses both the original data and the augmented data to train the model, thus verifying the superiority of our approach and the effectiveness of the data augmentation method. The second set of experiments uses the enhanced data on a single stream model using only channel domain information or spatial domain information.

3.2.1. Experiment 1

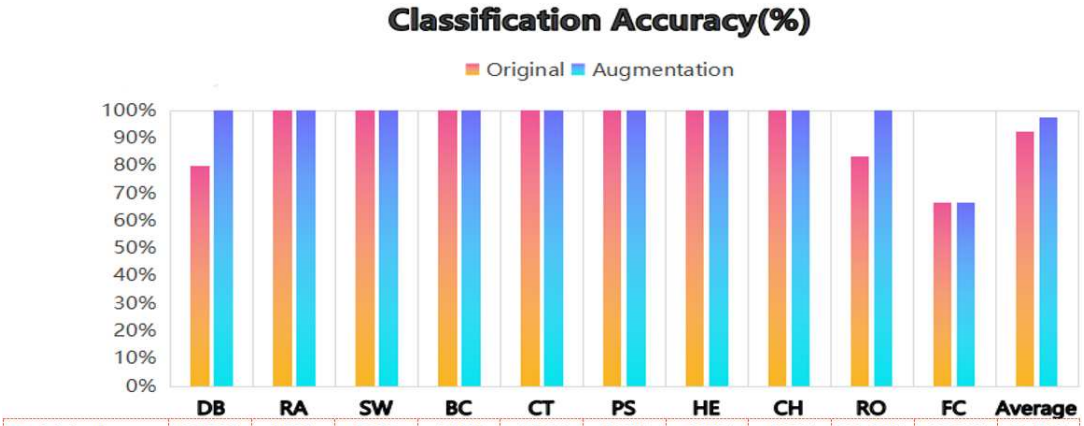
The superiority of our model compared to other baseline models and the effectiveness of the data augmentation methods used in this paper.
The purpose of this experiment is to compare the classification results of our proposed method with those of the deep learning methods proposed in recent years. Compared with single feature models: Piczak (2015) Piczak-CNN, Tokozume et al. (2017) Env Net v2, Guzhov et al. (2021) ESResNet, Mu et al. (2021) TFCNN, Jangid and Nagpal (2022) Env-Resnet, Fang et al. (2022) Fast-ACNN all used a single feature representation. For Piczak-CNN, ESResNet and TFCNN used a two-dimensional feature map as input to extract deep features in a way similar to image classification tasks. Among them Piczak-CNN has numerous parameters and is not very accurate in classification. Env Net v2 used the original audio signal as input and achieves good classification results. ESResNet introduced residual networks as well as attention mechanisms to audio classification with good results. However, this method does not released its parametric model number, so it can only be compared to this method in terms of classification accuracy, the accuracy of the classification is superior to the proposed method without the extra dataset and slightly inferior to the method with the extra dataset, which is imagenet. TFCNN proposed temporal attention mechanism and frequency attention mechanism, these mechanisms enable better capture of time-frequency features and achieves a high accuracy at a low cost. Compared with multi-feature models: Zhang et al. (2018) VGG-like CNN, Su et al. (2019) TSCNN-DS,

and Liu et al. (2021) SC-DNN, Guo et al. (2022) TF-ATTENTION all used many types of feature representations. VGG-likes network architecture is similar to VGG, with the model input being a stack of mel spectrogram and gammatone spectrogram in depth, which improves the classification performance of the model by a more abundant feature representation. TSCNN-DS uses an ensemble learning approach, two different types of feature representations are fed into subnetworks for training, and finally the predictions of each subnetwork are ensemble by DS argument theory. SC-DNN connects feature maps from multiple Sub-DNN models for enhanced feature extraction. TSCNN-DS and SC-DNN both achieved high classification accuracy, but at the cost of many parameters. Compared with most of the methods described in the above-mentioned literature, the models proposed in this paper have achieved absolute improvements.
We compare our model with several outstanding ESC models above. The classification results and the total number of the parameters of several state-of-the-art approaches are shown in Table 3. The difference in size of our model on the two datasets above is due to the different sizes of the inputs.

Model	Repesentation	Feature	ESC-10	urbansound8k	Param	Extra data
Piczak-CNN (2015)	2D	Log-Mel-spec	90.20%	73.70%	26M	×
Env Net v2(2017)	1D	Raw	91.30%	78.30%	18M	×
VGG-like CNN + mix-up (2018)	2D	Mel-spec+GT-spec	91.70%	83.70%	1.12M	×
TSCNN-DS (2019)	2D+2D	Multiple Feature	-	97.20%	16.9M	×
ESResNet-Attention(2020)	2D	Log-power spec	94.25%	96.83%	-	×
	2D	Log-power spec	97.00%	98.84%	-	√
TFCNN(2021)	2D	Log-Mel-spec	84.4%	93.1%	1.6M	×
SC-DNN(2021)	2D	Multiple features	96.1%	98.1%	14.31M	×
Env-Resnet(2022)	2D	Mel-spec	-	96.76%	-	×
TF-ATTENTION(2022)	2D	Multiple features	97.25%	98.25%	-	×
Fast-ACNN(2022)	2D	Log-Mel-spec	94.75%	97.51%	-	×
Ours	2D	MFCC	97.50%	98.51%	1.71M(ubs8) 2.29M(ESC-10)	×

Table 3: Comparing different models on the assessed dataset.

Then, we will demonstrate the validity of the data enhancements used. The number of test datasets of ESC-10 is 40. Our classification results for the ESC-10 dataset and the confusion matrix are illustrated in Fig 3 , Fig 4. From Fig 3 and Fig Fig 4 it can be concluded that the proposed approach obtains a classification result of 97.5%. After data augmentation, the recognition accuracy was 100% for all categories except for FC, and the most difficult category to classify is FC.
The number of test datasets of UrbanSound8k is 873. Our classification results on the UrbanSound8k dataset and the confusion matrix are shown in Fig 5, Fig 6. From Figure 5 and Figure 6 it can be concluded that the proposed approach obtains a classification result of 98.51%. Except for CP, DR and SM categories, all other categories will have 98% or better ranking accuracy. It is notable that the CP, DR and SM categories still achieved ¿96% acceptable results.



	DB	RA	SW	BC	CT	PS	HE	CH	RO	FC	Average
Original	80.00%	100%	100%	100%	100%	100%	100%	100%	83.33%	66.66%	92.50%
Augmentation	100%	100%	100%	100%	100%	100%	100%	100%	100%	66.66%	97.50%

Figure 3: Per-class accuracy for the original ESC-10 database and the augmented dataset.

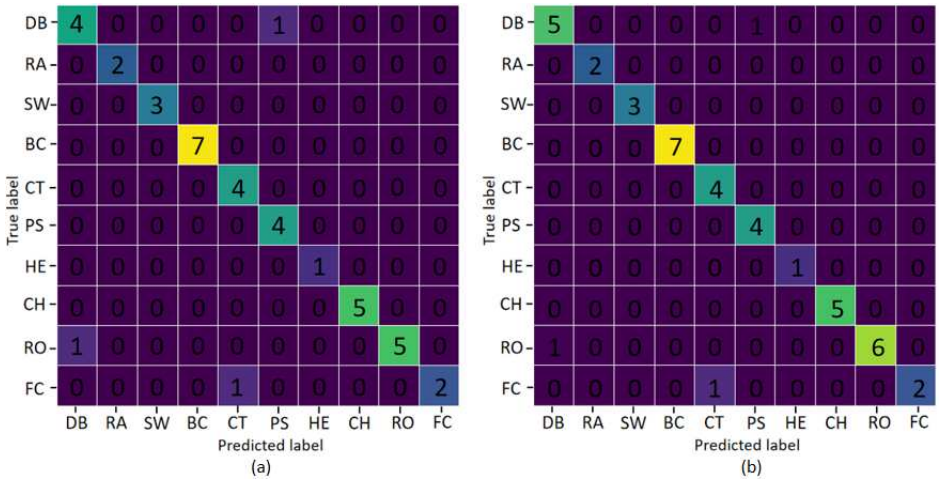
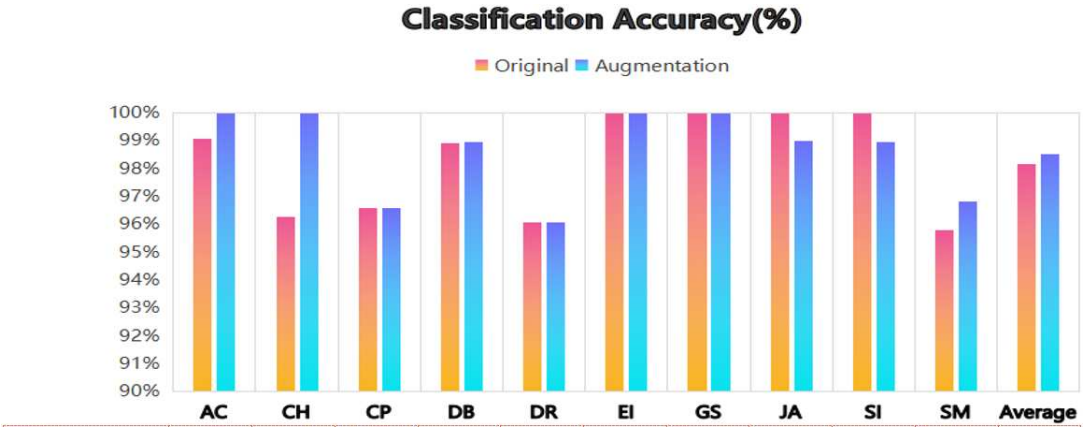


Figure 4: The confusion matrices on the ESC-10 dataset for the original data and the augmented data.(a)Original dataset.(b)Augmented dataset.



	AC	CH	CP	DB	DR	EI	GS	JA	SI	SM	Average
Original	99.08%	96.29%	96.59%	98.94%	96.08%	100%	100%	100%	100%	98.17%	98.17%
Augmentation	100%	100%	96.59%	98.59%	96.08%	100%	100%	99.00%	98.95%	96.84%	98.51%

Figure 5: Per-class accuracy for the original UrbanSound8K database and the augmented dataset.

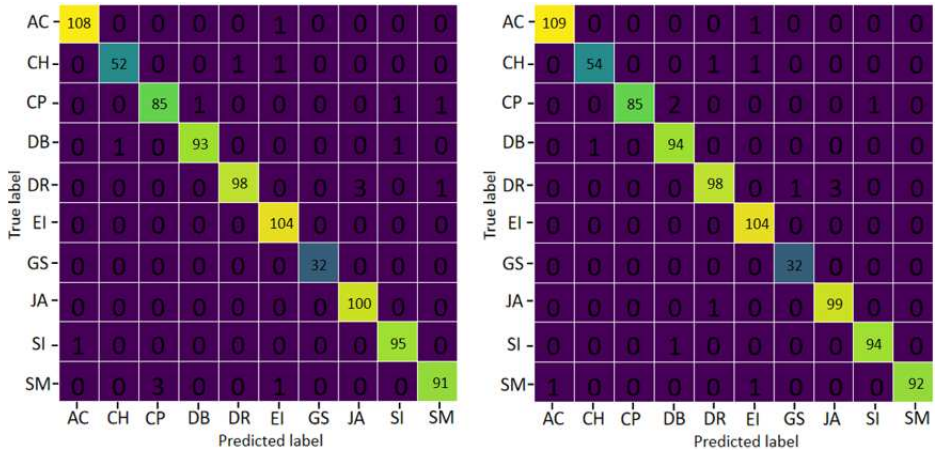


Figure 6: The confusion matrices on the UrbanSound8K dataset for the original data and the augmented data.(a)Original dataset.(b)Augmented dataset.

3.2.2. Experiment 2

Here, we will experimentally demonstrate that acceptable classification accuracy can be achieved by using our proposed convolutional block to classify only time-domain or frequency-domain features of the MFCC.
The experimental results are shown in Table 4. From Table 4, it can be concluded that classification is much more effective when time and frequency domains are combined than when just time or frequency domains are used. It can also be concluded that our network is more sensitive to frequency-domain features, and its classification is obviously more effective than classification using time-domain features. Fig 7 and 8 show the classification

Datasets	Loss	Domain	Accuracy	Param
ESC-10	0.4278	time	92.50%	1.68M
	0.6699	frequency	95.00%	608K
	0.3738	time&frequency	97.50%	2.3M
UrbanSound8K	0.1916	time	95.99%	1.1M
	0.1541	frequency	98.17%	608k
	0.0856	time&frequency	98.51%	1.7M

Table 4: Comparison of classification accuracy for time-domain features, frequency-domain features and combined time and frequency domain features.

confusion matrix for ESC-10 and UrbanSound8K using single domain information. From the classification confusion matrix using time-domain information and frequency-domain information, respectively, their error distributions differ significantly. The time-domain classification model is better at identifying audio signals with strong temporal sequences, but is weaker at classifying audio signals with more discrete distributions. The frequency-domain model is the exactly opposite. For example, the time-domain model in the ESC-10 dataset is less effective in classifying the DB (Dog Bark) and BC (Baby Crying) categories, which have relatively discrete sound signal distributions. The time-domain model is better for the time-series CT (Clock Tick) category, while the frequency-domain model is worse for this category. In the UrbanSound8K dataset, the frequency-domain model performs significantly better than the time-domain model in classifying the two more discrete categories of audio signals, CP (Children Playing), SM (Street Music).

4. Conclusion

In this paper, we proposed a two-stream convolution architecture based on audio feature disentanglement for ESC. Our first attempt to disentangle audio feature. In addition, we compared the computational complexity of our proposed approach and existing work. The experimental results of the ESC-10 and UrbanSound8K datasets demonstrated the efficacy of the proposed methodology and achieved advanced or competitive classification accuracy with low computational complexity. We also compared the impact of using individual information domains on the classification results. We conclude that either time domain or

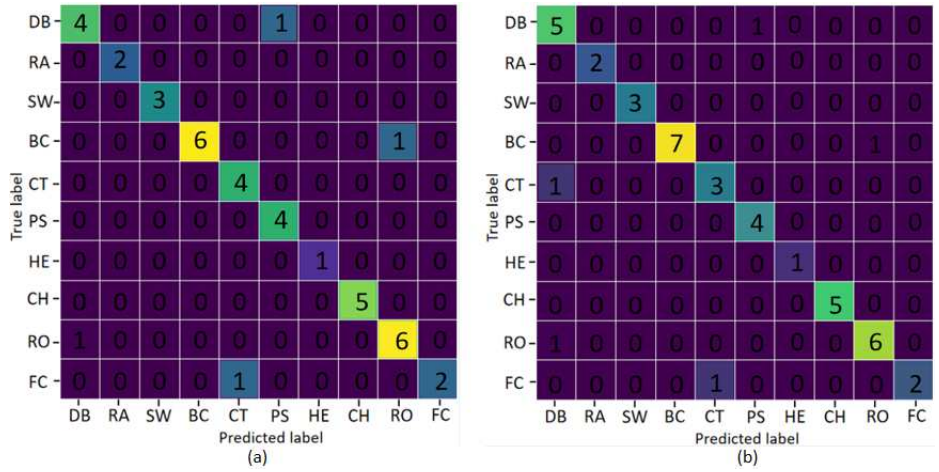


Figure 7: The confusion matrices on the ESC-10 dataset for the time-domain and the frequency-domain.(a)Time-domain.(b)Frequency-domain.

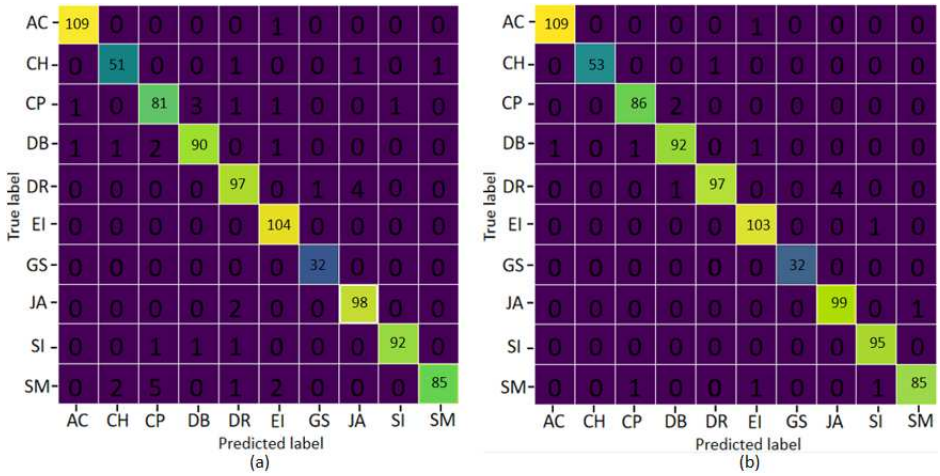


Figure 8: The confusion matrices on the UrbanSound8K dataset for the time-domain and the frequency-domain.(a)Time-domain.(b)Frequency- domain.

frequency domain information of audio features can be used as a basis for audio classification, but in our experiments, the use of frequency domain information clearly leads to better classification results. Experimental results show that the proposed method achieves superior classification accuracy with fewer parameters. In the future, we will further improve the robustness and accuracy of our model by using additional datasets such as Audioset to pre-train our model.

References

Osama Abdeljaber, Onur Avci, Serkan Kiranyaz, Moncef Gabbouj, and Daniel J Inman. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *Journal of Sound and Vibration*, 388:154–170, 2017. URL https://sciencedirect.53yu.com/science/article/abs/pii/S0022460X16306204.

Sajjad Abdoli, Patrick Cardinal, and Alessandro Lameiras Koerich. End-to-end environmental sound classification using a 1d convolutional neural network. *Expert Systems with Applications*, 136:252–263, 2019. URL https://arxiv.53yu.com/pdf/1904.08990.

Xiaochun An, Frank K Soong, and Lei Xie. Disentangling style and speaker attributes for tts style transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:646–658, 2022. URL https://arxiv.53yu.com/pdf/2201.09472.

Onur Avci, Osama Abdeljaber, Serkan Kiranyaz, and Daniel Inman. Structural damage detection in real time: implementation of 1d convolutional neural networks for shm applications. In *Structural Health Monitoring & Damage Detection, Volume 7*, pages 49–54. Springer, 2017. URL https://link.springer.com/chapter/10.1007/978-3-319-54109-9_6.

Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3):16–34, 2015. URL https://arxiv.53yu.com/pdf/1411.3715.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. URL https://arxiv.53yu.com/pdf/1206.5538).

Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan. Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis. *arXiv preprint arXiv:1904.02373*, 2019. URL https://arxiv.53yu.com/pdf/1904.02373.

Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 421–425. IEEE, 2017. URL https://arxiv.53yu.com/pdf/1610.00087.

Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980. URL https://ieeexplore.ieee.org/abstract/document/1163420.

Xifeng Dong, Bo Yin, Yanping Cong, Zehua Du, and Xianqing Huang. Environment sound event classification with a two-stream convolutional neural network. *IEEE Access*, 8: 125714–125721, 2020. URL https://ieeexplore.ieee.org/iel7/6287639/8948470/09136659.pdf.

Zheng Fang, Bo Yin, Zehua Du, and Xianqing Huang. Fast environmental sound classification based on resource adaptive convolutional neural network. *Scientific Reports*, 12(1): 1–18, 2022. URL https://wwwnature.53yu.com/articles/s41598-022-10382-x.

Jinming Guo, Chuankun Li, Zepeng Sun, Jian Li, and Pan Wang. A deep attention model for environmental sound classification from multi-feature data. *Applied Sciences*, 12(12): 5988, 2022. URL https://www.mdpi.com/2076-3417/12/12/5988/htm.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresnet: Environmental sound classification based on visual domain models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4933–4940. IEEE, 2021. URL https://arxiv.53yu.com/pdf/2004.07301.

Mahesh Jangid and Kabir Nagpal. Sound classification using residual convolutional network. In *Data Engineering for Smart Systems*, pages 245–254. Springer, 2022. URL https://linkspringer.53yu.com/chapter/10.1007/978-981-16-2641-8_23.

Aditya Khamparia, Deepak Gupta, Nhu Gia Nguyen, Ashish Khanna, Babita Pandey, and Prayag Tiwari. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*, 7:7717–7727, 2019. URL https://ieeexplore.ieee.org/iel7/6287639/8600701/08605515.pdf.

Serkan Kiranyaz, Adel Gastli, Lazhar Ben-Brahim, Nasser Al-Emadi, and Moncef Gabbouj. Real-time fault detection and identification for mmc using 1-d convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 66(11):8760–8771, 2018. URL https://ieeexplore.ieee.org/abstract/document/8353773.

Shaobo Li, Yong Yao, Jie Hu, Guokai Liu, Xuemei Yao, and Jianjun Hu. An ensemble stacked convolutional neural network model for environmental event sound recognition. *Applied Sciences*, 8(7):1152, 2018. URL https://www.mdpi.com/2076-3417/8/7/1152/pdf.

Minkyu Lim, Donghyun Lee, Hosung Park, Yoseb Kang, Junseok Oh, Jeong-Sik Park, Gil-Jin Jang, and Ji-Hwan Kim. Convolutional neural network based audio event classification. *KSII Transactions on Internet and Information Systems (TIIS)*, 12(6):2748–2760, 2018. URL https://www.koreascience.or.kr/article/JAKO201821464986105.pdf.

Chengwei Liu, Feng Hong, Haihong Feng, Yushuang Zhai, and Youyuan Chen. Environmental sound classification based on stacked concatenated dnn using aggregated features. *Journal of Signal Processing Systems*, 93(11):1287–1299, 2021. URL https://www.researchgate.net/publication/354617299_Environmental_Sound_Classification_Based_on_Stacked_Concatenated_DNN_using_Aggregated_Features.

Rui Lu, Zhiyao Duan, and Changshui Zhang. Metric learning based data augmentation for environmental sound classification. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2017. URL https://labsites.rochester.edu/air/publications/lu2017metric.pdf.

Wenjie Mu, Bo Yin, Xianqing Huang, Jiali Xu, and Zehua Du. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11(1):1–14, 2021. URL https://wwwnature.53yu.com/articles/s41598-021-01045-4.

Zohaib Mushtaq and Shun-Feng Su. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167:107389, 2020. URL https://www.sciencedirect.com/science/article/pii/S0003682X2030493X.

Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015. URL https://asset-pdf.scinapse.io/prod/1972567154/1972567154.pdf.

Yu Su, Ke Zhang, Jingyu Wang, and Kurosh Madani. Environment sound classification using a two-stream cnn based on decision-level fusion. *Sensors*, 19(7):1733, 2019. URL https://www.mdpi.com/1424-8220/19/7/1733/pdf.

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*, 2017. URL https://arxiv.53yu.com/pdf/1711.10282.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021. URL https://proceedings.neurips.cc/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019. URL https://arxiv.53yu.com/pdf/1812.04342.

Zhichao Zhang, Shugong Xu, Shan Cao, and Shunqing Zhang. Deep convolutional neural network with mixup for environmental sound classification. In *Chinese conference on pattern recognition and computer vision (prcv)*, pages 356–367. Springer, 2018. URL https://arxiv.53yu.com/pdf/1808.08405.

Zhichao Zhang, Shugong Xu, Shunqing Zhang, Tianhao Qiao, and Shan Cao. Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing*, 453:896–903, 2021. URL https://sciencedirect.53yu.com/science/article/pii/S0925231220313618.