IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 October 2022, accepted 12 November 2022, date of publication 16 November 2022,
date of current version 22 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3222495

RESEARCH ARTICLE

Research on Scattering Transform of Urban Sound Events Detection Based on Self-Attention Mechanism

SHEN SONG[1], CONG ZHANG[2], AND ZHIHUI WEI[1]
[1]School of Mathematics and Computer, Wuhan Polytechnic University, Wuhan 430048, China
[2]School of Electrical and Electronic Engineering, Wuhan Polytechnic University, Wuhan 430048, China

Corresponding author: Cong Zhang (hb_wh_zc@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61272278; in part by the Natural Science Foundation of Hubei Province under Grant 2015CFA061, Grant 2020CFB761, and Grant 2018CFB408; in part by the Hubei Provincial Major Science and Technology Special Projects under Grant 2018ABA099; and in part by the Hubei Provincial Department of Education Research under Grant D 20201601.

ABSTRACT Urban sound event detection can automatically preload relevant information for a robot to ensure that it can be applied to various scene-activity tasks. To address the limitations of timbre similarity and scene recognition by audio collection devices, a fusion model based on the self-attention mechanism is proposed in this paper. The model consists of scattering transform and self-attention model. The scattering transform computes modulation spectrum coefficients of multiple orders through cascades of wavelet convolutions and modulus operators. It is learnable compared with Mel-scale Frequency Cepstral Coefficients (MFCC), and can be used to better restore the semantic features of some sound scenes with similar timbres. The transformer has an outstanding effect on Natural Language Processing (NLP) owing to its self-attention mechanism. In this paper, the self-attention mechanism in its encoder was used in the model, mainly to make the feature granularity consistent to refine the features. In addition, Focal Loss function was adopted in the model to curb the sample distribution imbalance. The Google Command and ESC-50 were used to supplement the scene categories of dataset UrbanSound8K. The model parameters of the learnable filters that performed well on the dataset UrbanSound8K were preserved to fine-tune the other two datasets with insufficient data volume and more target categories. The length of slice duration was further explored the in the model. The experimental results show that the model can achieve better performance in a large range of scene models.

INDEX TERMS Preload information, scattering transform, feature granularity consistency, self-attention mechanism, focal loss.

I. INTRODUCTION

Urban sound event detection has shown good application prospects in our daily life, such as monitoring patients in the hospital for possible falls, collisions or other abnormal sounds, and reminding nurses of responding in time [1], monitoring the sound events of stolen trees and mountain fire that may exist in the forest [2], emotional classification [3], machine damage detection [4], etc. Urban sound event detection can assist video surveillance, reduce the number of video surveillance devices, and solve the problems of video surveillance being affected by light, blind spots in video surveillance, and expensive surveillance devices. In machine intelligence [5], Urban sound event detection can automatically preload relevant information for a robot. However, there are also some problems with sound event detection, such as poor noise immunity [6], weak information-carrying capability of sound, poor multi-source sound recognition owing to waveform interference, weak recognition ability, similar timbres, and scene recognition limited by sound collection device limitation.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman.

Traditional handcrafted acoustic features (e.g.,Mel Filter banks) have certain limitations. The problem with this method is the Fourier transform, which makes it successful. The Fourier transform is represented by a series of triangular wave expansions on an orthogonal basis, which is a global orthonormal basis lacking localization ability and is quite sensitive to noise, thus the waveform through the Fourier transform is also called a sine wave. This Orthogonality is more convenient for calculating coefficients. However, its premise is that the signal is a representation of a smooth and stationary signal and the Fourier transform can achieve an approximate optimal representation. However, the signals encountered in daily life are often not smooth signals but signals with many singular points. However, the performance of the Fourier transform on singular point signals is poor. It must be approximated with a large number of triangular waves of different frequencies, and the calculation of coefficients is slow and complicated, resulting in slow audio processing and Gibbs effect [7]. However, the timbres are determined by the high-frequency distribution of the spectrum and amplitude of each frequency. To reduce the computation of digital signals, a high-frequency spectrogram is discarded by default during processing, which means that MFCC is difficult to distinguish the signals with similar timbres.

From the perspective of application scenarios, A major difficulty of audio representations for classification is the multiplicity of information at different time scales: pitch and timbre at the scale of milliseconds, the rhythm of speech and music at the scale of seconds, and the urban sound event over minutes and hours. Mel-frequency cepstral coefficients (MFCC) are efficient local descriptors at time scales up to 25 ms. Capturing larger structures up to 500 ms is however necessary in most sound scene.

From the perspective of spectrum, Spectrograms compute locally time-shifting invariant descriptors over durations limited by windows. High-frequency spectrogram coefficients are not stable to variability caused by time-warping deformations, which occur in most signals, particularly in audio. Stability means that small deformations in signals produce small modifications of the representation, measured with a Euclidean norm. It is particularly important for classification. Mel-frequency spectrograms are obtained by averaging spectrogram values over Mel-frequency bands. It improves stability to time-warping, but it also removes information. Over time intervals larger than 25 ms, the information loss becomes too important, which is why Mel-frequency spectrograms and MFCC are limited to short time intervals. Modulation spectrum decompositions characterize the temporal evolution of Mel-frequency spectrograms over larger time scales [8], with auto correlation or Fourier coefficients. However, this modulation spectrum [9] also suffers from instability to time-warping deformation, which degrades classification performance.

The scattering transform [10] builds invariant, stable, and informative signal representations for classification, which are computed through a cascade of wavelet transforms and modulus non-linearities to recover the lost information. As a result, the scattering coefficients can be calculated over larger window sizes without as great of a loss of information, allowing larger-scale structures to be captured. These larger-scale structures include timbral structures, such as attacks, amplitude and frequency modulations, and interference phenomena found in musical chords. It is stable to deformations, which makes it particularly effective for image, audio and texture discrimination. The computational structure was similar to a convolutional deep neural network. It outputs time-averaged coefficients, providing informative signal invariants over potentially large time scales. What's more, the scattering transform has striking similarities with physiological models of the cochlea and of the auditory pathway.

A. RELATED WORK

In response to the aforementioned problems of MFCC, a considerable amount of research has been proposed to address its shortcomings.

Victor [11] compared spectrograms decomposed by Principal Component Analysis (PCA), Independent Component Analysis (ICA), Decomposition Analysis (FA) and Non-negative Matrix Factorization (Convolutive NMF) on large ASC datasets map dictionaries and spectrums of different sizes. It was shown that there is a correlation between different dictionaries and the size of the spectral feature map. Abidin [12] adopted Constant-Q Transform (CQT) for the audio signal, and adopted Local Binary Patterns (LBP) to extract its texture features from the transformed time-frequency signals, which are fed to the model of random forest for importance identification. However, when the scale of the spectrum changes, the encoding of the LBP features will be incorrect, and the LBP features will not be able to correctly reflect the texture information. In complex environments, the recognition effect is significantly reduced. Zhao Ren [13] stated that wavelet transform is not necessary, and fused scalograms (bump and morse) and spectrograms which are more suitable for ASC tasks, as they represent the signal in detail. However, Different scales are suitable for various tasks. It is necessary to select an appropriate scale for features extraction based on this task. In addition to the bump and morse scalograms, there are, for example, the Bark scale and Equivalent Rectangular Bandwidth. These scales suit different corresponding sound events but not ASC task. Geiger [14] proposed adopting Gabor Filter banks features to detect target events in different noisy background scenes. In the detection of non-stationary sound events, the implementation shows that Gabor features have better detection and classification performance than MFCC. But it is not suitable for multi-scale audio.

The above studies all try to moderate the drawbacks MFCC, whose Mel scale is concentrated in the low-frequency part and are sparse in the high-frequency part. It is unsuitable for multi-scale audio. Moreover, MFCC cannot detect target events in noisy background scenes. STFT inevitably leads to

the loss of formant. The problem of similarity of timbre was not addressed. Therefore, some researchers tend to extract features from raw audio to retain its spectrum as much as possible, but not through Fourier transform. CNNs are the most popular architecture for processing raw speech samples, because weight sharing, local filters, and pooling help discover robust and invariant representations.

Palaz D [15] tried to model the original signal directly and used a ''convolution-maximum-pooling-convolution'' model structure instead of MFCC to achieve the extraction of short-time features. It was shown that these features are susceptible to noise. Exploiting the parallel between time and frequency-domain processing is optional to improve robustness. Wei Dai [16] indicate that 2-layered CNNs are insufficient to extract discriminative features from raw waveforms for sound recognition at the front end. This is in contrast to models using the spectrogram as input, which achieves good-performance with only just two convolutional layers. The receptive field of the first layer 320 layers was down to 80, and the model accuracy increased by 6.6%. However, the small RF model has many more dispersed bands, and thus a lower frequency resolution for subsequent layers. Conversely, the large RFmodel has fine-grained filters, but does not have sufficient filters in the high-frequency range, showing that it cannot effectively respond to local high-frequency impulses. Hoshen [17] presented a DNN architecture for speech acoustic modeling from multichannel waveforms, which can reduce the noise level and improving recognition performance compared to Mel-fb magnitude-based baseline. With the network filter length, pooling window and hop chosen to match a Mel-fb baseline, the model learns a bank of bandpass beamformers that qualitatively follow an auditory filterbank-like scale and has spatial selectivity that exploits the structure of the data. However, Traditional CNN kernel filters are not efficient at learning common acoustic features because of the lack of constraints on the neural parameters. Ravanelli [18] proposed SincNet, a neural architecture for directly processing waveform audio, inspired by the way filtering is conducted in digital signal processing, which imposes constraints on the filter shapes through efficient parameterization. Beyond improvements, SincNet also significantly improves the convergence speed over a standard CNN and is more computationally efficient for exploitation of filter symmetry. An analysis of the SincNet filters reveals that the learned filter bank is tuned to precisely extract some important characteristics, such as pitch and formants. However, the low and high cut-off frequencies are the only parameters of the filter learned from the data. This solution still offers considerable flexibility, but does not force the network to focus on high-level tunable parameters with a broad impact on the shape and bandwidth of the resulting filter. Gauthier [19] proposed complex gabor-based SincNet on a phoneme recognition task, which is an optimal time-frequency resolution alternative to the SincNet architecture. It is shown that the proposed approach can produce results comparable to those of state-of-the-art systems while operating on a raw waveform.

Some researchers have concluded that features such as MFCC are more suitable for feature extraction, and that the experimental results depend on the ability of the classifier. The works of the classifier.

The researches have been addressed with features such as MFCC and classifiers based on GMMs, XGBoost or SVMs [20], [21], [22]. Other approaches use some form of DNN, including CNNs [23], RNNs [24], and CRNNs [25]. With the emergence of ResNet and attention mechanism, related models [26], [27] have been applied in various fields. Jianyu L [28] proposed a multi-scale convolutional capsule network (MCCN), integrating low-level and high-level features in a convolutional neural network (CNN) as multiscale features are conducive to noise reduction and robust feature extraction, and a capsule network (CapsNet) is used to recognize the spatial relationships in attitude data. Kong et al. [29] proposed the structure of Wavegram-Log-Mel- CNN to train pretrained audio neural networks (PANNs) on large-scale audio datasets and convolved the convolved features from the original wave graph. Concatenate with the Log-Mel transformed channel features.

Since the representation of the spectrogram is the frequency band under the time frame, multi-scale feature extraction can alleviate the problem of feature inconsistency under different tasks. However, the front-end model is more important for restoring the features and reducing the feature loss in the extraction to restore the volume between the formant frequencies in the timbre.

B. CONTRIBUTIONS

In this paper, we propose a fusion model based on a self-attention mechanism to restore the semantic features of sound scenes with similar timbres and make the feature granularity consistent to refine the features. Our contributions can be summarized as follows:

We explore a scattering transform that consists of learnable filters, which can better deal with the brown noise widespread in urban sound events, and better restore some semantic features of sound scenes with similar timbres, and its DNN structure can better filter noise, providing good feature support for subsequent feature recognition. The model focuses on the impact of each time frame early in the 1D convolution like Squeeze-and-Excitation network, but drops the Squeeze-and-Excitation structure. This is because recognition of the model is related to the weight of each bin. Therefore, there was no Squeeze-and-Excitation process related to the channel. In addition to the noise influence, it is necessary to perform Gaussian filtering on the weights to reduce the influence of noise on the weights and the impact of spectral loss replacing global pooling. The structure can also be used to compress invalid ones and enhance important time frames. Using this method, the number of network layers can be reduced to prevent overfitting. The bandwidth and center frequency in the filters can be learned and adjusted according to the task, which is adaptive.

The filtered features are fed into a self-attention network. This was used to adjust the time-frequency resolution. The self-attention mechanism in the model is adopted to refine the features by keeping the feature granularity consistent, and it can obtain its global features at the early stage, allowing the model to achieve better recognition results.

With the above methods, the self-attention mechanism model proposed in this paper for sound event detection is well trained. However, UrbanSound8k has only 10 low-level categories, which cannot well generalize all urban sound event types well. Therefore, the categories of Google Command, ESC-50, etc. were supplemented after the UrbanSound8K dataset, and the effectiveness of the model was verified. The research also fine-tunes the learnable filter using a transfer learning method and takes the parameters on the Mel filters as the initial parameters of the learnable filter on the UrbanSound8K dataset. However, the ESC-50 and Google-command datasets consists of insufficient data. We freeze the first multi-head attention block of its pretrained model and retrain the previous filter layers. Because the similarity of the dataset was poor, it was important to retrain the higher layers and filters based on the dataset. The experimental results demonstrate that the model can accurately detect sound events. Compared with other classical residual networks and the networks with the ''Squeeze-and-Excitation'' mechanism in the classifier, the model proposed in this paper shows better performance. This improves the effectiveness of sound event detection.

II. METHODS

A. LEARNABLE FILTERS

The front ends can be categorized according to the procedures they perform. There are two key categories: scattering transform (FST) [30] based front ends and Short-Time Fourier Transform (STFT) based front ends. Unlike STFT, which multiplies the filter banks matrix with the spectrogram, FST adopts a convolutional layer on the raw audio waveform to approximate a standard filtering process. FST-based front-ends methods have made considerable progress, scattering transform can learn the relationship between harmonics to realize the effective detection of sound events, which can extract the reverberation and phase information to summarize the speech signal. Compared to the Mel spectrum, it loses less information and achieves a better detection effect.

Moreover, Many STFT-based front-ends are fixed and may not be well-suited for certain downstream tasks. Both types of front ends employ a filter-like transform to simulate the non-linear sensitivity of the human ear to frequency. The distribution of filter center frequencies is called scale. Mel scale can capture human perception of pitch relatively well. There are also lesser-known Bark and Equivalent Rectangular Bandwidth (ERB) scales. However, these ratios are mostly based on past experiences and are fixed equations. To make such operations in the front end domain adaptive, filters can be made learnable. Filter banks can learn its center frequency and bandwidth.

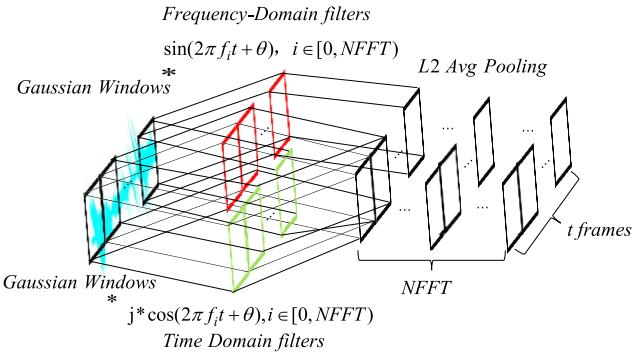


FIGURE 1. Scattering transform.

As shown in Figure 1, scattering transform representation process is that after the signal passes through the Gaussian window, it will be sent to the Gabor Filter banks [31]. The Filter banks can learn the center frequency f_i and bandwidth a of each filter through backpropagation, and they are all constrained and learnable parameters. The audio signal passes through $NFFT$ filters in the time and frequency domain respectively, to form 2^*NFFT time domain signal and frequency-domain signal bins, and then squares the frequency domain signal bins corresponding to the time domain, getting its Hilbert Spectrum. Determine the information of time domain bins in each channel according to the out-channel numbers.

Complex Gabor filters are defined as the product of a Gaussian kernel multiplied by a complex sine function, as shown in Equation (1):

$$g(t) = ke^{j\theta}w(at)s(t) \tag{1}$$

where $w(at)$ and $s(t)$ are as in Equation (2), (3)

$$w(at) = e^{-\pi(at)^2} \tag{2}$$

$$s(t) = e^{j(2\pi f_0 t)} \tag{3}$$

$g(t)$ can be decomposed into frequency-domain and time-domain signals, and the process is shown in Equation (4):

$$\begin{aligned} &ke^{j\theta}s(t)e^{j(2\pi f_0 t+\theta)} \\ &\quad = k(\sin(2\pi f_0 t+\theta), j\cos(2\pi f_0 t+\theta)) \end{aligned} \tag{4}$$

where k, θ, f_0 denotes the filter parameters. The complex Gabor filters can be considered as two out-of-phase filters, in the real and complex parts of the complex function, respectively.

The real-part Gabor filters representation is shown in Equation (5), which performs a sinusoid transform after the Gaussian kernel.

$$g_r(t) = w(t)\sin(2\pi f_0 t+\theta) \tag{5}$$

The imaginary Gabor filters representation is shown in Equation (6), which performs a cosine transform after the Gaussian kernel.

$$g_i(t) = w(t)\cos(2\pi f_0 t+\theta) \tag{6}$$

The real and imaginary components of a complex Gabor filter are phase sensitive, this is their response to a sinusoid is another sinusoid. By obtaining the magnitude of the output (square root of the sum of squared real and imaginary outputs) we can obtain a response that is phase insensitive and thus an unmodulated positive response to a target sinusoid input. In certain cases, it is useful to compute the overall output of the two out-of-phase filters. One common way to do so is to add the squared output (the energy) of each filter; equivalently, we can obtain the magnitude. This corresponds to the magnitude (more precisely the squared magnitude) of the complex Gabor filter output. In the frequency domain, the magnitude of the response to a particular frequency is simply the magnitude of the complex Fourier transform, i.e.

$$\|g(f)\| = \frac{k}{a}\hat{w}(\frac{f-f_0}{a}) \tag{7}$$

This is a Gaussian function centered on f_0, with a bandwidth proportional to a. Therefore, the center frequency response of the filter is f_0, and in order to obtain the full width at half maximum (FWHM, half-magnitude), the calculation is shown in Equation (8).

$$\hat{w}(\frac{f-f_0}{a})=\mathrm{e}^{-\pi\frac{f-f_0}{a^2}} = 0.5 \tag{8}$$

The bandwidth obtained by transform is $0.46797a$, which is about $0.5a$. The calculation process is shown in Equation (9).

$$f - f_0 \pm \sqrt{a^2 \log 2\pi} = 0.4697a \approx 0.5a \tag{9}$$

The learnable bandwidth is strictly constrained between $-a\sqrt{2\log 2\pi}$ and $a\sqrt{2\log 2\pi}$, and the center frequency is constrained between $-1/2$ and $1/2$.

And the f_{max}, f_{min} of each filter is initialized by Mel scale, and the signal is constrained within Mel scale, firstly is converted to Mel scale. As shown in Equation (10):

$$mel\ scale = 2595\log(1+\frac{f}{700}) \tag{10}$$

The obtained center frequency and bandwidth equally divided on the Mel scale were converted into frequencies, as shown in Equation (11). to obtain the initialized center frequency and bandwidth with the Mel scale of the frequency.

$$f = 700^*(10^{\frac{mel}{2595}} - 1) \tag{11}$$

The filtered signals obtained from the Gabor filter layer and square mode layer were the Hilbert envelope. The envelope is then sent to several layers of one-dimensional convolution, which adds an extra branch to the shortcut connection. The shortcut connection is adopted to solve the problem of deep neural network degradation, and Dilated convolution is used to reduce the number of network layers.

The overall front-ends structure is shown in Figure 2, whose first layer adopts a scattering transform based on a constrained learnable Gabor filter. The following down sampling layers imitated the shortcut connection of ResNet by adopting a Gaussian filter in its branch. The reason for adopting Gaussian filter is that, after obtaining the Hilbert envelope, the signal output has the same time resolution as the input, which need to be down-sampled to a lower sampling rate to obtain valid information. However, direct convolution or 2D convolution will result in the need for a deeper network to obtain a sufficiently large receptive field, but a deeper network will lead to a decrease in the recognition. This problem can be solved with methods like max pooling or average pooling, but there are better ways to do so. Zhang [32] showed that in standard 2D convolutional architectures, including ResNet [33] and DenseNet [34], replacing max pooling and average pooling layers with (fixed) low-pass filters can improve the performance of image classification. In feature extraction, we employ a single shared low-pass filter for all frames, but we implement low-pass filtering by depth wise convolution such that each kernel is associated with a low-pass filter. Each kernel in the learnable front-ends have a different bandwidth and center frequency, and a specific lowpass filter can be learned for each kernel. Furthermore, compared with the pooling methods, low-pass filtering can weaken the details, noise, edges and sudden changes in the audio, which is shown in Equation (12), is obvious in data compression and noise reduction. The bandwidth and center frequency in per low-pass filter function can be learned, initialized with a bandwidth of 0.4, resulting in a frequency response close to the Hann window used by the Mel Filter banks. In order to enable the feature extraction system to fully extract the global features of the audio instead of localized features, a 12-layer 1D convolution is used to express the high-level semantics of the obtained envelopes.

$$\phi_n(t) = \frac{1}{\sqrt{2\pi}a}e^{-\frac{(t-f_0)^2}{2a}}, t = -\frac{a}{2}, \ldots, \frac{a}{2} \tag{12}$$

While its weight is multiplied by each filter, and the center frequency and bandwidth parameters of each filter are sent to the back-propagation network for learning. The Squeeze-and-Excitation structure was dropped, but with a full connection layer. The task is not in the shape of (b, c, h, w), shaped in (b, f, t). In addition, the Squeeze-and-Excitation structure, which focuses on the channel, was applied to its channel. However, our task focused on filters.

B. SELF-ATTENTION MECHANISM

After obtaining the input that preserves its sequence information from the 1D convolution of learnable filters. The purpose of the multilayer scattering structure is to reduce spectrally unnecessary signals it conveyed by it. In contrast to the vit Transformer [35] in computer vision, the input is divided into different patches for normalization, and then a linear layer is applied to each patch to reduce the dimension and embed the position information, and then sent to the Transformer model, avoiding the explosion of pixel-level self-attention block's operation. The learnable filter can reduce the signal loss while expanding the receptive field after the scattering transform,

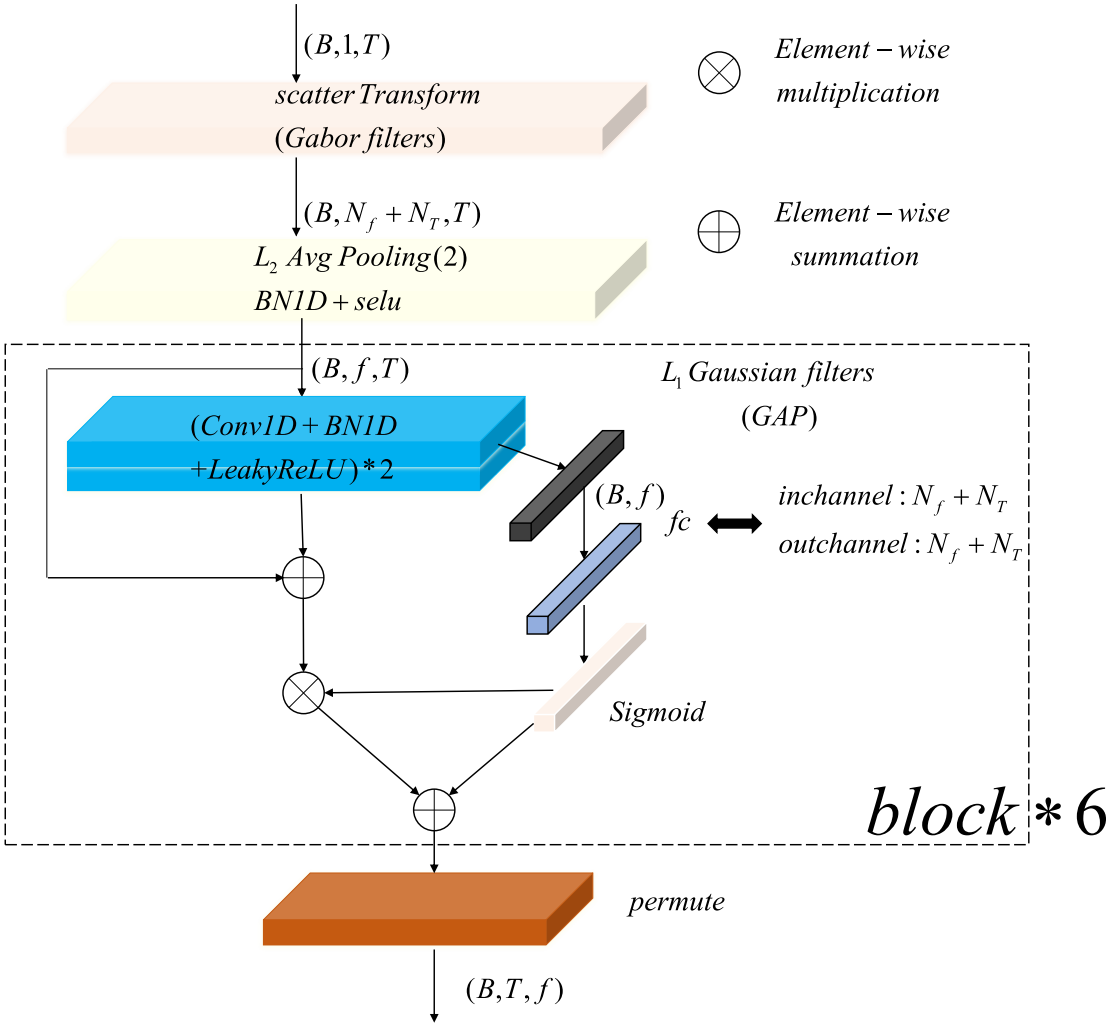


FIGURE 2. Learnable filters network with Scattering transform.

and avoid the explosion of the operation in the spectrum while saving the time series.

Convolutional network models such as ResNet are good at identifying texture features, but ignore their expressions for detailed features. Previous models relied heavily on convolution to model correlations between different regions. The convolution operator has only a local receptive field, and the long-range correlation can only be post-processed by several subsequent convolution layers. The expression for the correlation cannot be represented by small convolutions. Correlation optimization algorithms may have difficulty coordinating multiple layers to capture these correlated parameter values, and when these parameters are applied to the validation set, the accuracy and generalization ability of the model will decrease. Increasing the size of the convolution kernels can increase the representational ability of the network, but in the meanwhile, it also losses the computational and statistical efficiency gained by using local convolutional structures. This demonstrates a better balance between the ability to model long-range dependencies and the computational and statistical efficiency. The self-attention module computes the response of a location as the weighted sum of all the location features, where the weights (or attention vectors) are computationally inexpensive. The self-attention module computes the response at a certain location as a weighted sum of all the location features, whose weights (or attention vectors) are computationally inexpensive. Convolution processes information in local neighborhoods, and using convolutional layers alone is computationally inefficient for modeling the long-range dependencies of features. Attention mechanisms have become an integral part of models that capture global correlations.

The standard Transformer [36] accepts a sequence of 1D token embeddings as the input. In this paper, in order to deal with the two-dimensional spectrogram, the token embedding operation is not performed, and in the positional embedding stage, the *d_model* is replaced with *nfft*, where is the resolution of the spectrogram. The classification head is

implemented by a hidden layer during pre-training and a single linear layer during fine-tuning, which contains two nonlinear GELU layers. The layer norm (LN) was used before each block, and a residual connection was used in each block. This model is shown in Figure 3.

The spectrogram of 1D convolution forms the input sequence. When feeding a higher-resolution spectrogram, this yields a larger effective sequence length while keeping the time series size constant.

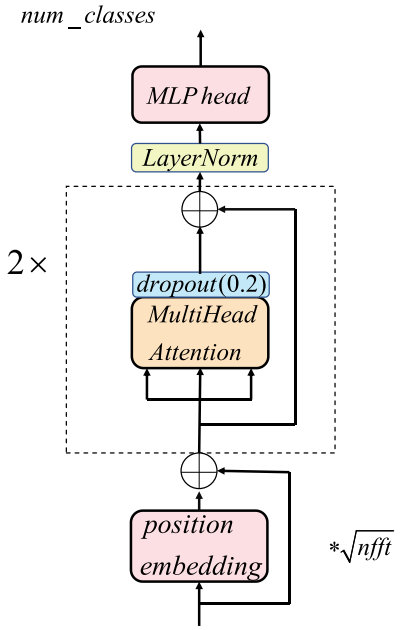


FIGURE 3. Self-attention mechanism model.

The representation of multi-Head attention is shown in Figure 4. The correlation score between each frame in the time series needs to be obtained, and the correlation score can be calculated by using the dot product method, which is to calculate the dot product with each vector in Q and each vector in K, Vector Q and vector K are both filtered and compressed sequence signal in the model. The matrix corresponding to the correlation score is: $score = QK^T$. The score is a matrix in the shape of (T, T). Subsequently, the score of the correlation between each frame in the input sequence is normalized, and the purpose of normalization is mainly to stabilize the gradient during training. $score = score/\sqrt{d_k}$, d_k is the dimension of vector K. Using the *soft* max function, the score vector in each frame is converted into a probability a distribution in [0, 1], highlighting the relationship in its time frames. Multiply the probability distribution in the frames by the corresponding Value, $Z = softmax(score)V$, V is shaped in $(T, nfft)$, $(T, T) \times (T, nfft)$ gets the final matrix Z shaped in $(T, nfft)$. The overall calculation is shown in Equation (13):

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{13}$$

On the basis of this self-attention mechanism, multi-Head Attention only uses one set of the input embedding matrix W^Q, W^K, W^V to transform to obtain Query, Keys, Values, and then each group is calculated to obtain a matrix Z Finally, the obtained multiple Z matrices are concatenated. The Multi-Head Matrix of 8 group are used in the model. After getting the matrix Z through multi-Head Attention, it is not directly passed to the fully connected neural network FNN.

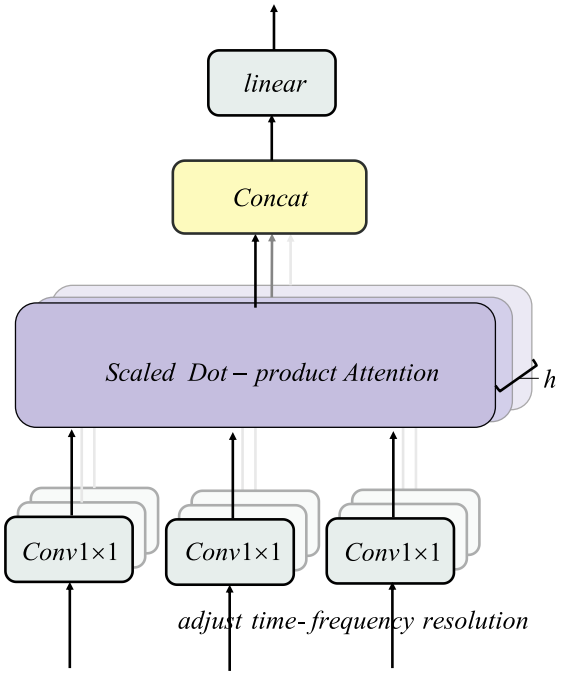


FIGURE 4. Multi-head attention.

Because the 12-layer one-dimensional convolution in the early stage can fully obtain audio features, a linear layer is not required to extract the features; therefore, the linear is removed. Scaled dot-product attention is adopted to adjust the time-frequency resolution to achieve the consistency of feature granularity.

Besides, it's a problem of unbalanced sample categories, and the cross-entropy loss function cannot solve this problem very well. Therefore, a more comprehensive Focal Loss method [37] was adopted. which is similar to the case of channel attention, and a function was used to measure the total loss of difficult and easier-to-classify samples. Depending on the difficulty of the classification, the weight of the easier-to-classify samples is reduced, allowing the model to focus more on difficult-to-classify situations during training. Its operation is given in Equation (14).

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{14}$$

Compared with the cross-entropy function, Focal Loss has a modulating factor $(1 - p_t)^\gamma$. For accurately classified samples $p_t \to 1$, the modulating factor approaches zero. For the inaccurately classified samples, the modulating factor is up to 1. That is, compared with the cross-entropy loss function, Focal Loss function does not change the loss for samples with

inaccurate classification, and the loss decreases for samples with accurate classification. Overall, this is equivalent to increasing the weight of the inaccurate samples in the loss function. This also reflects the difficulty of classification. The larger the value, the higher the confidence of the classification and the easier it is for the representative sample to be divided. Therefore, Focal Loss is equivalent to increasing the weight of difficult samples in the loss function, making the loss function tend to be difficult samples, which helps improve the accuracy of difficult samples.

C. NETWORK-BASED DEEP TRANSFER LEARNING

The dataset ESC-50 has sufficient data categories, per which there is little data. This is the primary reason that transfer learning is applied to the other two datasets. The general network is a model that obtains the hierarchical feature representation of data through pre-training, and then uses high-level semantic classification. The bottom layer of the model contains low-level semantic features (for example, edge information, color information, etc.), which are actually invariant in different classification tasks, and the real difference is the high-level features. Transferring features from distant tasks may be better than using random features. Usually, the first several layers are not particularly related to the specific image dataset, and the last layers of the network are closely related to the selected dataset and its task objectives. The first several layer features are called general features in the article (general) features, and the last several layers are called specific features.

Network-based deep transfer learning [38] refers to reusing part of the pre-trained network in the original domain, including its network structure and connection parameters, and transforming it into a part of the deep neural network for the target domain. First, the network was a source-domain trained using a large-scale training dataset. Second, part of the network preprocessed in the source domain is transferred to a new network designed for the target domain. Finally, the fine-tuning policy can be updated for the transmitted subnetworks. Training deep learning models from scratch based on small samples is difficult because a large number of weight parameters must be adjusted, which are generally randomly initialized.

Transfer learning [39] has potential to overcome the above-mentioned problems by reasonably applying the existing knowledge gained from related but different domains. Various transfer learning strategies have been applied to solve several pattern recognition problems. Parameter transfer, the most widely applied transfer learning strategy, is not only easier to implement, but also more suitable for classification tasks with auxiliary training data.

There are only 2000 pieces of data in the dataset ESC-50, but there are 50 classes, each with only 40 pieces of data. When split in a ratio of 8:2, the data for each category is unbalanced and the amount of data is small. The model trained on the dataset of UrbanSound8K is partially transferred to the network designed in the target domain, and the first layer of unimportant extraction of edge, texture information and power information in the self-attention mechanism is frozen for the transmitted sub-network. However, the extraction process of the filter is related to a specific scene; therefore, it is necessary to load the pre-trained model of the transferred filter model, and adjust its center frequency, weight, bandwidth, and other parameters according to the training data. The layers close to the MLP of the self-attention network extract a high-level semantic feature representation. It is also necessary to preload training according to the task, of realizing a fine-tuning strategy for a network with insufficient data. The MLP layers of the network are closely related to the selected dataset and its task objectives, which cannot be frozen and must be trained with the data.

III. EXPERIMENTS AND DISCUSSION

A. EXPERIMENTAL DATASET

The urban sound events listed in Table 1 contain four main categories: human, natural, mechanical, and music. UrbanSound8K [40], provided by DCASE, contains 10 low-level categories of urban sounds: air conditioners, car horns, children playing, dog bark, drilling, engine idling, gunshots, hand hammers, sirens, and street music. Except for children playing and gunfire, all other categories were selected because of their high frequency in urban noise complaints. However, they cannot represent all environmental classes. Therefore, ESC-50 [41] provided by Kaggle and Google Command [42] provided by Google were added to the categories of the experiments. Google Command mainly supplements the speech in the categories, ESC-50 mainly supplements categories such as Movement, Plants, and Non-motorized Transport in the Table 1.

In order to prevent feature differences caused by inconsistent feature granularity, all audios were uniformly resampled and sampled to 44.1 KHz, then converted to mono, and then clipped subsequently, and time offset was applied to move the audio to the left or right. The random amount is shifted to the right to augment the original audio signal, and finally obtain raw-audio. and sent to the network to generate the spectral envelope. The training and validation of the model were split in a ratio of 8:2.

B. THE PERFORMANCE OF LEARNABLE FILTERS

The most common audio feature is the Mel-scale Frequency Cepstral Coefficients (MFCC). The MFCC features extracted from raw audio were compared with the method of extracting envelope features using the scattering transform proposed in this paper. A comparison of the features is shown in figure 5. The audio features only take the data of one batch, and the boxplots are compared among the first 16 channels. It is obvious that the learned channel features can converge to the range in $-0.6745\sigma \sim 0.6745\sigma$. Several obvious problems can be observed in the figure 5. The features represented by MFCC are more likely to contain noise, and feature extraction is more chaotic. The features learned by

TABLE 1. The detailed categories in the urban sound events.

	Categories					
Urban Acoustic Environment	Human	Voice	Speech			
			Crying			
			…			
			Children			
		Movement	Footsteps			
	Nature	Elements	Wind			
			Water			
			Thunder			
		Animals	Bark			
			Howl			
			Tweet			
		Plants	Leaves			
	Mechanical	Construction	Jackhammer			
			…			
			Hammering			
			Drilling			
		Ventilation	Air Conditioner			
		Nonmotorized Transport	Bicycle	Spokes		
				Bell		
			Skateboard			
		Social/Signals	Bells			
			Alarm			
			Gunshots			
		Motorized Transport	Marine			
			Rail	Train		
				Subway		
			Road	Car	Police	Siren
					Ambulance	
				Motorcycle	Police	
					Private	
				Bus		
				Truck	Fire Engine	
					Garbage Truck	Siren
			Airplane Helicopter			
	Music					

constrained convolution can learn more expressive features and significantly suppress noise. Because it learns the center frequency and bandwidth, the median behaves differently in position, and the center frequency behaves differently. The bandwidth is limited between $Minimum(Q_1 - 1.5 * IQR)$ and $Maximum(Q_3 + 1.5 * IQR)$. The center frequency is within IQR, where represents the distance between the third quartile and the first quartile (Interquartile Range). It can be clearly seen that the center frequency and bandwidth changed. However, owing to the constraints, the center

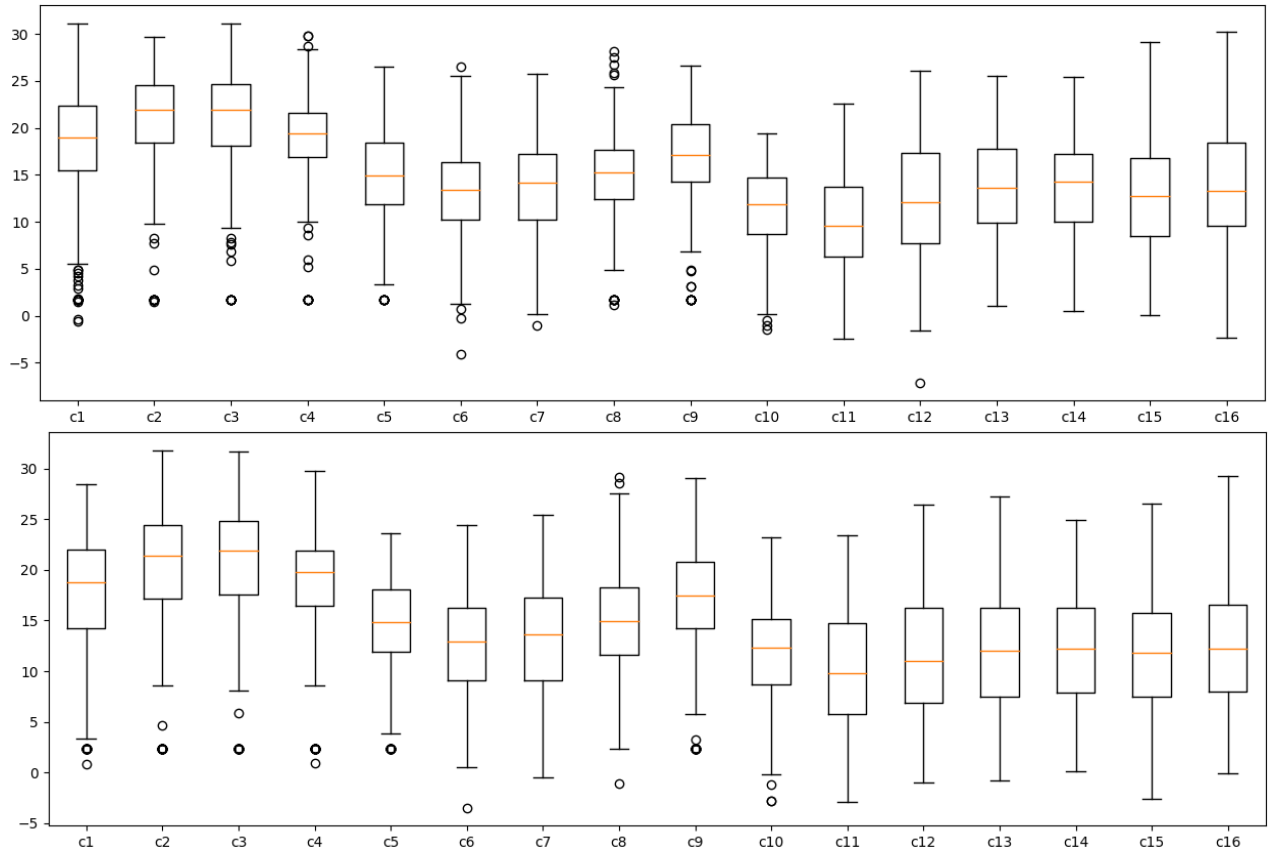


FIGURE 5. In a batch: Upper is the feature boxplots obtained by MFCC, and lower is the feature boxplots after 10 epoch scattering transform.

frequency and bandwidth of its learning will not deviate significantly.

C. EXPERIMENTS SETTING

The experiment was conducted in an Ubuntu16.04 operating system, and the framework of audition, pytorch and Lingvo was applied in the experiments. IZotope Radius is selected in the audition to stretch the audio and pitch simultaneously. To reduce the influence of artifacts on features, this study adopts a high time-frequency resolution and sets nfft to 1024. Each model uses 16 audio data as a batch, initializes the learning rate to 0.001, window of kernel size to 1024, and hop size to 320 samples and uses the Adam optimizer to iteratively update the parameters. Adam can dynamically adjust the learning rate so that the learning rate is closer to the current state of parameter update, so that the model can converge better, as shown in Table 2.

For the other systems based on log Mel spectrograms, STFT was applied to the waveforms with a Hamming window of size 1024 and a hop size of 320 samples. This configuration resulted in 100 fps. We used 64 Mel filter banks to calculate the log Mel spectrogram. The lower cut-off frequencies of the Mel banks were set to 50 Hz to remove low frequency noise. We use torchlibrosa, a PyTorch implementation of functions of librosa to build log Mel spectrogram extraction into models.

TABLE 2. Optimizer and stratagegy.

Category	Value
Optimizer	Adam
Learning rate	1e-3
Max rate	1e-2
batch	16
Epoch	100
nfft	1024
Kernel size	1024
Hop size	320
num_blocks	2
d_model	1024
attention_heads	4

D. MAXIMUM SLICE DURATION ON THE MODEL

After comparing the features extracted by the learnable filter with the MFCC. The discussion results are shown in Figure 6, and it can be observed that the maximum slice duration is better between 4-6 seconds relatively. Therefore, we selected 5s as the maximum slice duration for each audio. In the first experiment, we investigated how the choice of threshold affects the performance of the model. To do this, we generated ten copies audio in UrbanSound8K, and the maximum slice duration for each copy was changed from 10s to 1s. To ensure that the observed variation in accuracy was not an artifact of a particular classification algorithm, we compared

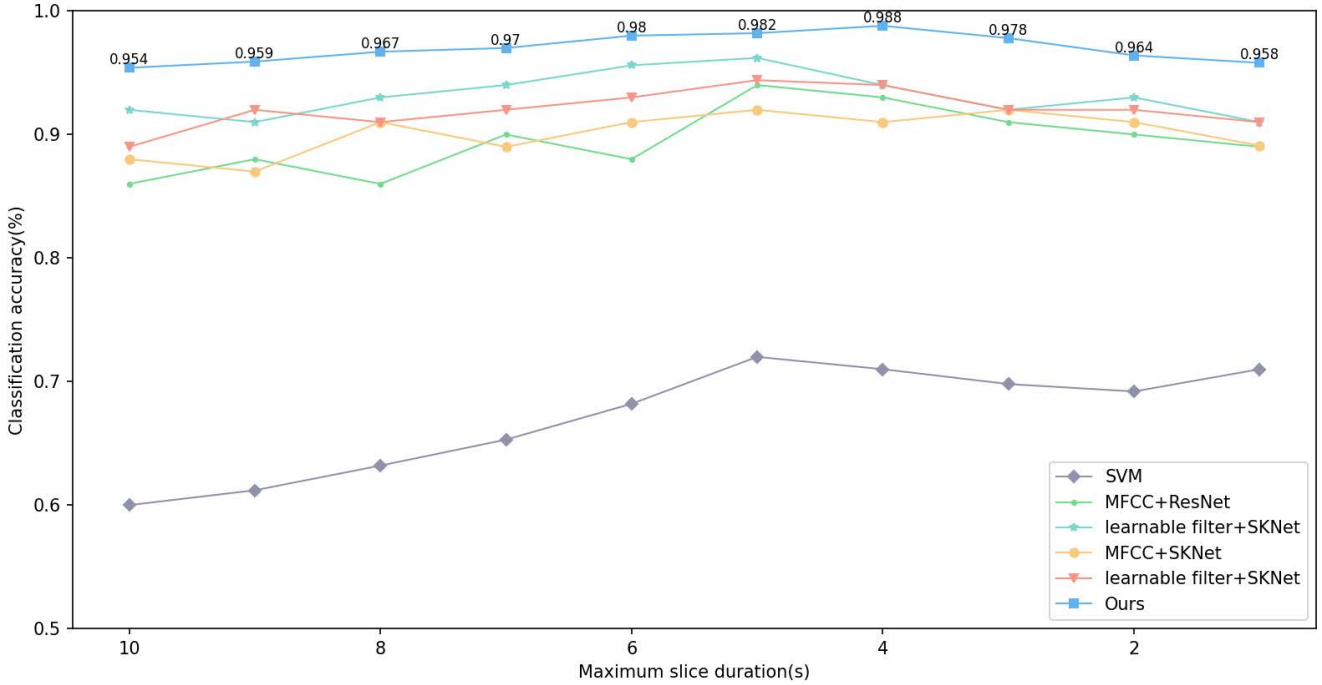


FIGURE 6. The performance of learnable filters combined with classical models for maximum slice duration.

six combined front-ends + classifier algorithms: MFCC + ResNet, learnable filter + ResNet, MFCC + SKNet, learnable filter + SKNet, support vector machine (radial basis function kernel), and the learnable filter-self-attention model adopted in this study. The traditional method was found to perform poorly in practice. The MFCC under the same model and parameters compared with scattering transform, the performance of which will still be significantly degraded. Because there is no backpropagation process in SVM, it is meaningless to use a learnable filter; therefore, there is no comparison between MFCC and scattering transform.

The results show that we observe consistent behavior for all classifiers except MFCC + ResNet: the performance remains stable from 10s to 6s, after which it starts to gradually decrease. Consider the best performing classifier (Ours), with no statistically significant difference between performance using 6s slices and 4s slices (whereas below 4s, the difference becomes significant), and choose 4s slices.

Figure 7 shows that different sound categories are affected differently by maximum slice duration: categories such as car horn and drill have fast events that are clearly identifiable on short time scales and are therefore largely unaffected by duration; whereas street music, siren and children Play etc. decreased almost monotonically, but this shows the importance of analyzing these courses on longer time scales, and suggests that multiscale analysis may be a relevant avenue for research. To understand the relative difference in performance between the classes, we examined the confusion matrix of our classifier on UrbanSound8K as shown. We found that the classifier tended to confuse three broad categories of air conditioners and idling engines, jackhammers and drills, children playing and street music. This is because the timbre of each pair is very similar (for the last pair, harmonics are a possible cause). To a certain extent, the influence of harmonics still exists and cannot be completely solved. However, the model in this study confirms that the harmonics can be identified.

E. EVALUATION INDICATORS

Several commonly used evaluation metrics are used in this study: precision, recall, F1 score, and confusion matrix.

Precision is the ratio of the number of correct predictions to all test samples. Its calculation formula can be expressed as (15):

$$precision = \frac{TP}{TP + FP} = \frac{\sum_{i=0}^{c} p_{ij}}{\sum_{i=0}^{c} \sum_{j=0}^{c} p_{ij}} \tag{15}$$

p_{ii} indicates that the prediction is class i, which is actually class i, and p_{ij} indicates that the prediction is class i, and the actual class is class j. Precision is represented by PRE, which represents the proportion of the correct audio category prediction to the total audio frequency, which can reflect the accuracy of the model classification to a certain extent.

Recall refers to the ratio of the number of correct predictions to all real results. The calculation formula is expressed as (16):

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

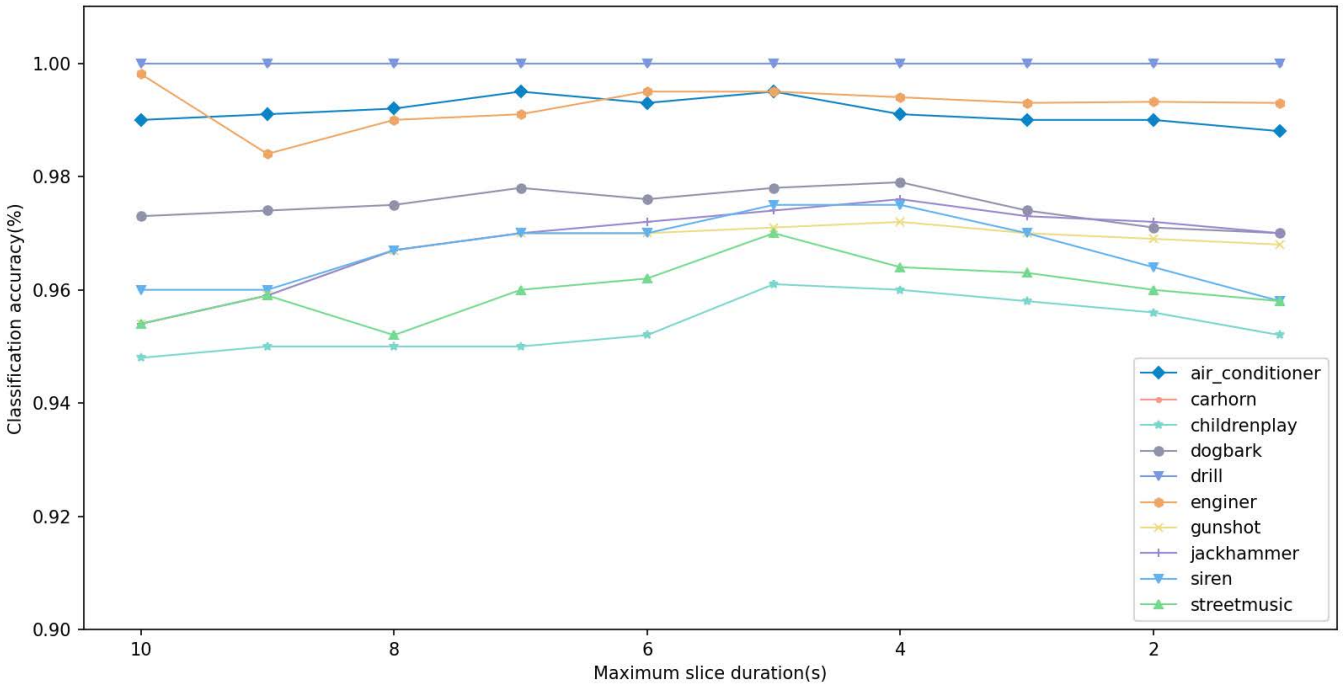


FIGURE 7. The performance of different categories on the maximum slice duration under our model.

TP: Predict True samples as True;
FN: falsely predict True samples as False;
FP: Predict False samples as True.

Macro-F1 Score: Also known as Balanced Score, it is defined as the harmonic mean of precision and recall. After calculating each class PRE and REC, calculate F1, and finally average F1. Its calculation formula can be expressed as (17):

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precison + recall} \tag{17}$$

Confusion matrix: An analysis table that summarizes the prediction results of the classification model and the records in the dataset in matrix form according to the two criteria of the real category and the category predicted by the classification model. where the rows of the matrix represent the true values and the columns of matrix represent the predicted values. That allows us to intuitively understand which kind of samples the model does not perform well.

The model in this study performed 100 iterations on the training data, and finally reached convergence. The precision reached 98.8% on the UrbanSound8k dataset and reached 96.7% and 87.32% for Google-Commands and ESC-50, respectively. Noise with different signal-to-noise ratios was added to the three data sets: 20 dB, 10 dB, and 0 dB. The performance is presented in Table 2.

From Table 3, it can be concluded that the MFCC front-ends are more sensitive to noise than the learnable front-ends scattering transform in this study, and the PRE of the acoustic event is reduced by 2%∼5% under various signal-to-noise ratios. The scattering transform is not very sensitive to noise performance, and the PRE of sound events is reduced by 0%∼1% under different SNR noises. As the self-attention mechanism can obtain global information at an early stage, its model can identify features at an early stage. Compared with the "Squeeze-and-Excitation" SKNet [43] model which obtains global information at later stage, this effect can be improved. for better recognition. The learnable filter is similar to the noise reduction structure of DNN, which can achieve a good noise reduction effect, and the combination of the two achieves a relatively good recognition effect.

Under the same SNRs, the values of the Precision and F1 score achieved by the scattering transform were constently higher than that obtained by MFCC. In addition, the lower the SNRs are, the larger the improvements obtained by Ours model are. For example, when the SNR is 0 dB, the scattering transform achieves a slight decrease of approximately 1% compared with 2∼5% under 0 dB in the same classifier. If scattering transform is adopted, better noise immunity can be achieved under different noise conditions.

As far as three individual classifiers are concerned, the effect of our model is better than that of the other two classifiers, whereas SKNet is the worst in terms of both Accuracy Recall and F1 score under different SNRs. Similar results were obtained for the other two datasets. However, it is intriguing that SKNet, as an attention mechanism for modeling between channels, has a significantly lower PRE than ResNet in terms of the recognition effect. We conclude that the early features extracted by MFCC are rather confusing, resulting in the inability to effectively identify key features during the learning process.

TABLE 3. Per-class performance of the sound event classification on urbansound 8K, google-CMD and ESC-50 datasets presented by mean value in percentage. pre: precision; rec: recall; F1: F-1 score.

Datasets	model	20dB			10dB			0dB		
		PRE	REC	F1	PRE	REC	F1	PRE	REC	F1
UrbanSound8k	MFCC + ResNet	0.94	0.92	0.931	0.92	0.91	0.913	0.902	0.901	0.9
	MFCC + SKNet	0.92	0.89	0.91	0.89	0.887	0.891	0.882	0.877	0.879
	Scatter + ResNet	0.962	0.954	0.955	0.957	0.958	0.958	0.956	0.94	0.932
	Scatter + SKNet	0.93	0.912	0.89	0.927	0.925	0.923	0.926	0.927	0.925
	Ours	0.988	0.982	0.981	0.98	0.979	0.978	0.98	0.977	0.978
Google-CMD	MFCC + ResNet	0.944	0.943	0.941	0.921	0.919	0.918	0.92	0.917	0.914
	MFCC + SKNet	0.939	0.937	0.932	0.912	0.913	0.91	0.908	0.906	0.908
	Scatter + ResNet	0.952	0.95	0.951	0.95	0.948	0.945	0.946	0.947	0.944
	Scatter + SKNet	0.943	0.941	0.94	0.939	0.94	0.941	0.939	0.942	0.94
	Ours	0.967	0.968	0.968	0.963	0.961	0.96	0.958	0.957	0.95
ESC-50	MFCC + ResNet	0.855	0.854	0.856	0.831	0.829	0.827	0.828	0.821	0.819
	MFCC + SKNet	0.82	0.823	0.83	0.79	0.762	0.772	0.782	0.775	0.77
	Scatter + ResNet	0.87	0.869	0.868	0.867	0.864	0.86	0.862	0.852	0.85
	Scatter + SKNet	0.865	0.853	0.842	0.86	0.861	0.865	0.863	0.86	0.859
	Ours	0.915	0.893	0.891	0.89	0.891	0.889	0.892	0.887	0.885

After a follow-up investigation, it was found that ResNet can achieve the best recognition effect at the 18th layer, whereas the channel attention of SKNet and a deeper neural network will lead to overfitting. This study finds that using two layers of the self-attention mechanism can achieve the best recognition effect and prevent overfitting. This can also explain why the SKNet performance has a 2%-4% accuracy gap compared to ResNet. For the ESC-50 dataset with a small amount of data and an uneven distribution of species, better results can be obtained. Compared with cross Entropy Loss [44], Focal Loss can achieve an improvement of 2%. The main basis was derived from an analysis of the data categories of the dataset.

Its confusion matrix on the UrbanSound8K dataset. The classification situation of the model in each category is shown more clearly. As can be seen from the figure, the model has a very high accuracy rate for the vast majority of the categories. 100% accuracy on driving and car h. However, there are 5 misjudgments in child and street, although in the case of a large cardinality of 204, the accuracy rate reaches 97.5%, and the degree of confusion is the highest in the entire audio classification. Followed by child and dog misjudgments each of 3, the data is second in the error, listen carefully to the audio, and find that the audio is mixed with the sound of the dog. Consequently, the learned features cannot be correctly distinguished. This is mainly because of the similarity of the scenes in which the difficulty of distinguishing increases. However, this did not affect its overall performance.

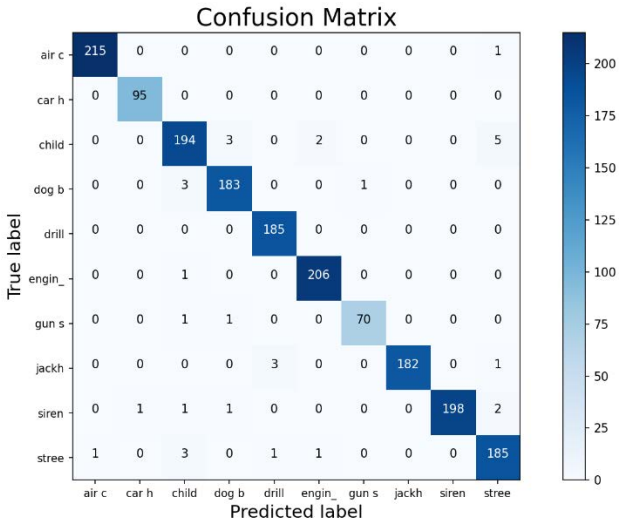


FIGURE 8. Confusion matrix on UrbanSound8K dataset.

Table 4 compares the results obtained in various recent studies with our model on the datasets of UrbanSound8K, ESC-50 and Google-Command. The results show that the proposed model marginally outperformed the state-of-the-art performance.

TABLE 4. Per-class performance of the sound event classification on urbansound 8K, google-CMD and ESC-50 datasets presented by mean value in percentage. pre: precision; rec: recall; F1: F1 score.

	Front-end	Improved(front)	Aggregation	Loss	Strategy	Acc (%)
UrbanSound8K						
Salamon J [45]	Log-Mel	8-frame patches of PCA	spherical k-means	MSE		73.7
Karol J [46]	Log-Mel		80 filters in each layer (CNN)	CEL		73.1
Dai W [16]	18layers CNN		ResNet16	CEL		71.8
Salamon J [47]	Log-Mel	DRC	SB-CNN	CEL	10fold	79
Zhang [48]	Log-Mel	Delta (concat)	Dilated conv+ max pooling+2fc	-		81.9
Abdoli S [49]	7layers of 1D-conv+maxpooling	Gammatone filter bank	SB-CNN+3 fc layers	-	10fold+Fine-Tuning	89.0
Li [50]	Log-Mel (static)	Raw Net CNN	DS-CNN	-	10fold+ Loss-Level Fusion	92.7
Yu S [51]	MFCC-CST	LM-CST	Two-Stream CNN	-	10fold+ Decision-Level Fusion	94.9
Mushtaq Z [52]	Log-Mel	Distribution of Augmented Data (concat)	DCNN (without max pooling)	-	5fold	95.3
Ours	scattering transform	Constraint learnable filters + 18 layers of 1D conv	multi-Head attention of Transformer	FL		98.8
Google-command						
Sattler F [53]	Log-Mel	Sparse ternary compression	VGG11	MSE		85.46
Yifan [54]	Log-Mel		Conformer Transformer	CEL	Two branches merged	86.3
Yin B [55]	Log-Mel		SRNNs	-		92.2
Ours	scattering transform	learnable filters + 18 layers of 1D conv	multi-Head attention of Transformer	FL	Fine-Tuning	96.7
ESC-50						
Tokozume Y [56]	Log-Mel (static)	Delta (concat)	Env Net	-	5fold	68.3
Li [50]	Log-Mel	Raw Net CNN	DS-CNN	-	10fold+ Decision-Level Fusion	82.8
Zhang Z [57]	Log-Mel	GTs (mixup)	VGG10	-		83.9
Zhang Z [58]	Log-Mel (static)	Delta (concat)	ACRNN	-		86.3
Mushtaq Z [52]	Log-Mel	Distribution of Augmented Data (concat)	DCNN (without max pooling)	-	5fold	89.2
Ours	scattering transform	Gabor filters + 18 layers of 1D conv	multi-Head attention of Transformer	FL	Fine-Tuning	91.5

We compare our model with some existing methods. On the dataset UrbanSound8K, 9 methods are mainly listed. From the traditional machine learning approach of Salamon J [45] to the Decision-Level Fusion of Two-Stream CNN of Yu S [51]. It can be concluded that most researchers are trying to improve the effect of the model on the basis of Log-Mel and MFCC, which, to a certain extent, shows that these traditional methods are difficult to effectively perform competent on the task alone, and the characteristics need to be supplemented. The inputs to the network consist of time-frequency patches (TF patches) extracted from the log-scale Mel spectrogram representation of the audio signal, as well as chrominance, spectral contrast, and Tonnetz features, among others. Comparing the models of Salamon J [47]and Abdoli S [49], we can see that under the same classifier and 10-fold cross-validation strategy, the features learned by the strategy of fine-tunning and front-end in 1D conv are better than the front-end in Log-Mel and no fine-tunning strategy. The front-end of Mushtaq Z [52] is based on Log-Mel, which concatenate the enhanced data in parallel, whose classifier is a deep convolutional network (without max pooling). A precision of 95.3% was obtained. In contrast, the network of Zhang [48] also adopts Log-Mel, but only achieves 81.9%, whose classifier drop the max pooling. It can be observed that max pooling had a negative effect on the model. The model of Mushtaq Z [52] still performs well on the ESC-50 dataset. Experiments may attribute the model success to data augmentation. However, several other models used data augmentation to a certain extent, although not in parallel. It was shown that max pooling negatively affects the training effect of the model during the training process. Li [50] adopted the model of taking Log-Mel features recognition as the main stream, extracting features from the raw waveform as weights and adopting the strategy of Loss-Level Fusion to obtain better

TABLE 5. Statistics of accuracy achieved by the three models.

	Paired Samples Test								
	Paired Differences							Significant	
	Mean	Deviation	Standard error mean	Confidence Interval (95%) Lower	Upper	t	Dof	Tails 1	Tails 2
MFCC + ResNet	-0.11324	2.61154	0.07082	-0.25215	0.02568	-1.599	1745	0.055	0.110
MFCC + SKNet	-0.17574	3.17933	0.08621	-0.34486	-0.00661	-2.038	1745	0.021	0.042
Scatter + ResNet	-0.04124	0.94461	0.02261	-0.08558	0.00310	-1.824	1745	0.034	0.068
Scatter + SKNet	-0.05735	2.72825	0.07398	-0.20248	0.08777	-0.775	1745	0.219	0.438
Ours	-0.01088	0.67912	0.01625	-0.04276	0.02099	-0.670	1745	0.252	0.503

results. This can better show that features extracted from the raw waveform have a positive effect on the model. However, because the learnable filters are unconstrained used, the learning parameters are affected by noise. The learning of the bandwidth and center frequency in the filters is weird.

On the dataset Google-command, Models [53], [54], [55] have shown that speech has obvious requirements for the identification of timing signals. The model proposed by Yifan [54] outperformed the self-attention of the Transformer and Conformer owing to the addition of peak detection. This technique alleviates the problem of similar timbres, in which a multi-Gaussian surrogate gradient is used by its Grid search.

This phenomenon was also observed for the dataset ESC-50. For example, Zhang Z [58] and Tokozume Y [56] used the same front-end, and the ACRNN model with time-series recognition achieved better recognition than the Env Net of CNN. It can be shown that the time-series signal has a significant impact on feature recognition. The Transformer of the self-attention mechanism can better solve the problem of gradient disappearance and gradient explosion in the long sequence training process, and is more suitable for audio overfitting tasks.

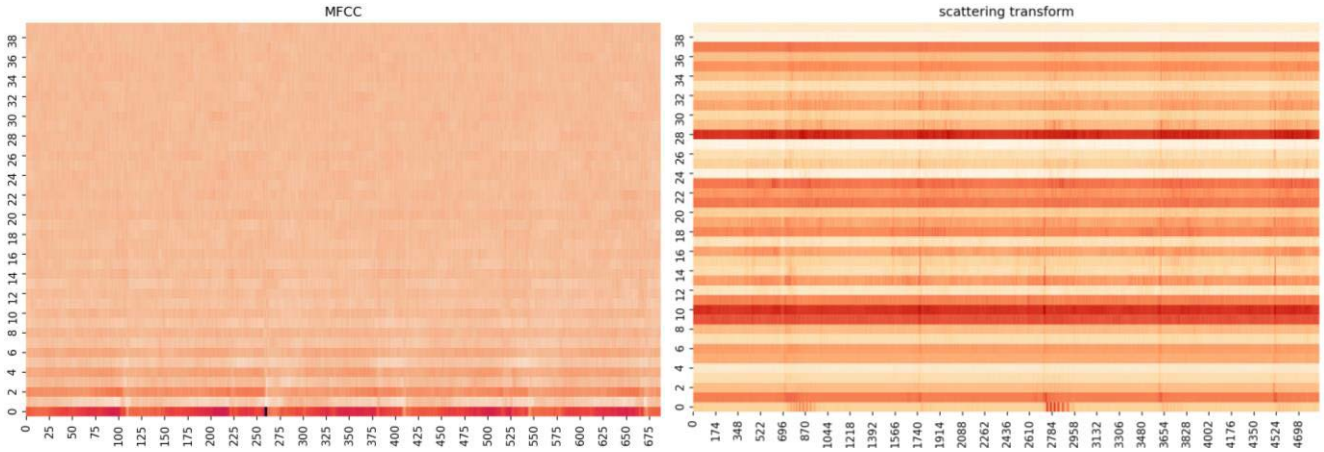
We conclude that max pooling affects the model more than the long-term dependency problem, in the case of insufficient data. The feature supplement of Log-Mel is an unavoidable problem for long-term sequences. The envelope feature cannot effectively cover all ranges and must be supplemented with peaks, pitch, and tonal space features. The main reason for this is that compression and Fourier transform truncate some unsolvable harmonics. This is also the problem our model tries to solve.

In order to further demonstrate whether our model outperforms the models which combined MFCC or scattering transform and other networks in a statistically significant way, we added the experimental results from paired accuracy statistics and applied a paired sample t-test. Table 5 shows the performance achieved by the five models with UrbanSound8K.

The standard error mean is obtained by taking the difference between the data generalized from the model and predicted data. If the model and solution space are the same, that is, $\mu_1 - \mu_2 = 0$ (as a known population mean μ_0). That is, the difference in paired data should fluctuate around 0 and not be too far away from 0, so this kind of data can be regarded as the sample mean of the difference. The represented unknown population mean μ_{dev} (Deviation)compared to the known population mean $\mu_0 = 0$.

The standard error mean obtained by our model is much smaller than that of the other models (0.01625), which proves that the effect of our model can reflect the solution space of the data. The number of Deviation and Mean in our model were smaller than those in the other composite models. It can be seen that adopting our model to search the solution space of the data is 0.03 higher than using the scatter + ResNet model, with a 95% CI of −0.04-0.02, and the difference was statistically significant ($t = -0.670, p > 0.05$). $p > 0.05$ proves that there is no significant difference between the predicted and actual data. While $p<0.05$ in the MFCC + SKNet Proves that there is a significant difference between the data predicted by the MFCC + SKNet model and the actual data. This can be explained by the fact that SKNet can easily lead to an overfitting state compared with ResNet. Scatter + SKNet can obtain prediction data that are not significantly different from the actual data, confirming that the features obtained by MFCC in the early stage are misleading, resulting in overfitting in the model learning process. The scattering transform can effectively extract these features.

We further show a visual feature map using a scattering transform and MFCC. Figure 9 shows the feature thermal map obtained by the scattering transform from learning the



(a) MFCC spectrogram (b) The heatmap of the signal in scattering transform

FIGURE 9. The visualization of MFCC and scattering transform.

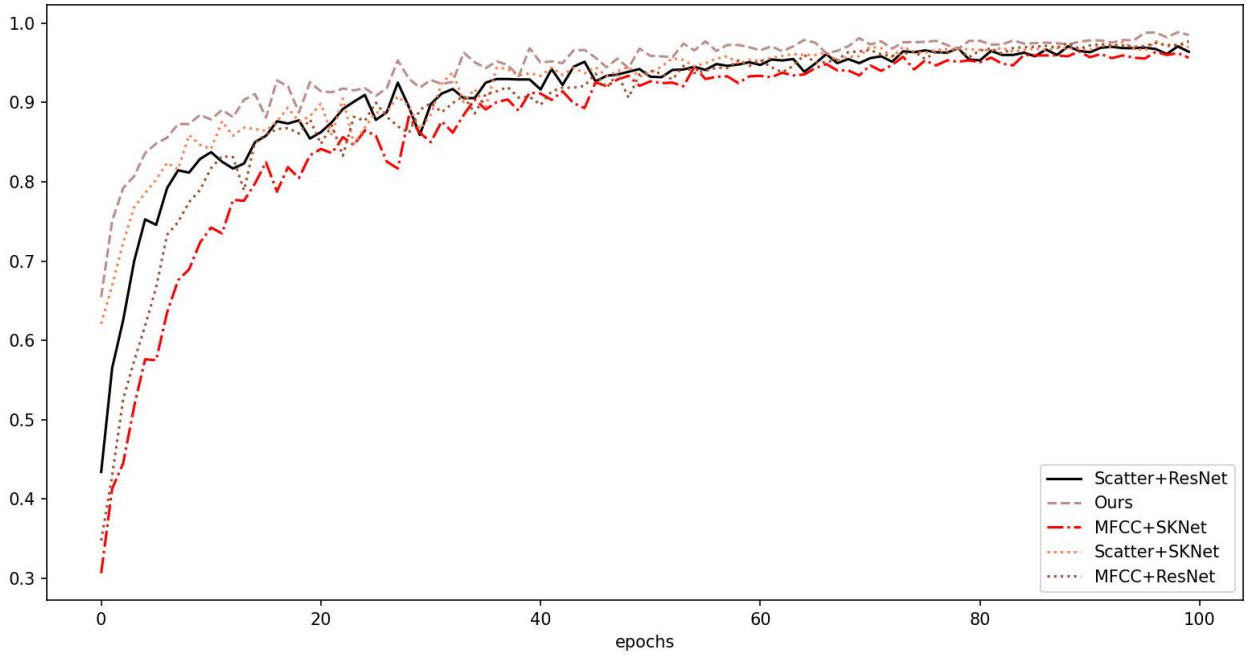


FIGURE 10. The mAP of the models during 100 epochs on the UrbanSound8K.

signal of gun shot on the dataset of UrbanSound8K, in which (a) is the MFCC spectrogram, (b) is the heatmap of the signal in the scattering transform. The light color is the background of the picture, and the darker color is the feature extracted by the model. The darker the color, the more important this feature is considered by the model. It is obvious that the front-end of scattering transform in this paper has obtained a more detailed feature map.

In addition, the research also compares the mAP of the models with self-attention mechanism and some mainstream models. The mAP of the epoch, as shown in Figure 10, is drawn, and the experimental results are listed in Table 3. It is obvious that our model can achieve better results at an early stage. The front-end of the scattering transform is generally better than the front end of MFCC.

Figure 11 shows the loss diagram of different models in the training process, where the abscissa is the number of iterations, and the ordinate is the loss value. It can be seen from the figure that the speed of the loss decline, whose rate indicates the speed of the model converges. The convergence speed of the scattering transform model is generally higher than that of the MFCC model.

Table 6 lists the accuracy of the fine-tuned our model. Our fine-tuned system achieved an accuracy of 0.915, outperforming previous state-of-the-art system. The Freeze front-end and Freeze_L2 systems achieve accuracies of 0.87

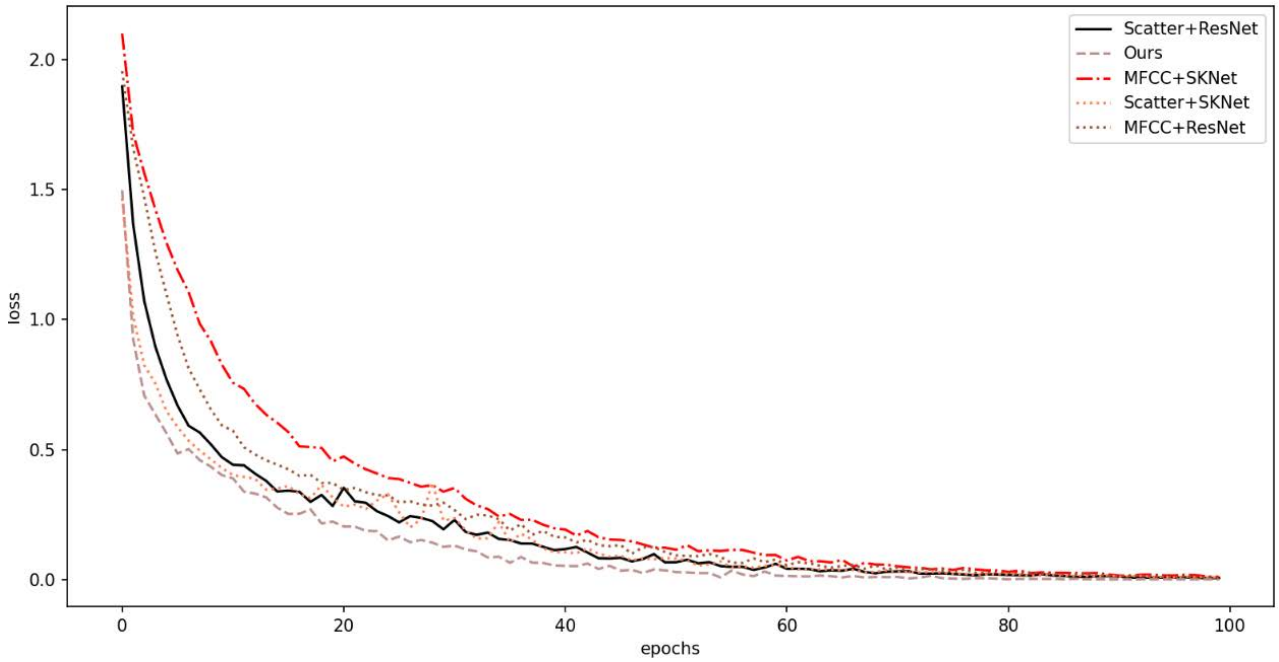


FIGURE 11. Variation of training losses.

TABLE 6. Accuracy of ESC-50 and google command.

	Scratch	Freeze front-end	Freeze_L1	Freeze_L2
Google Command	93.2	91.7	96.7	94.1
ESC-50	86.4	82	91.5	88.2

and 0.82, respectively. By contrast, training the system from scratch achieves an accuracy of 0.864. This phenomenon also exists in the Google Command. On the Google command, the fine-tuning effect is the best if the first layer of self-attention is frozen. In addition, if the second layer of self-attention is frozen, the effect is lower than that of freezing the first layer. If the front-end is frozen, it is not as effective as scratch. We can see that if we freeze the front end, the effect is even worse than that identified for the features extracted from the raw audio.

IV. CONCLUSION

In this study, a learnable self-attention model for sound event detection is proposed to alleviate the problem of inconsistent feature granularity caused by similar timbres and inconsistencies in collecting audio equipment. First, the fast Fourier transform was abandoned at the front-ends of feature extraction, and a learnable scattering transform was used. One-dimensional convolution is added to enhance its receptive field whereas imitating the residual block structure of ResNet, and Gaussian filtering is used on its shortcut branch. The filter performs feature filtering, and its structure can achieve the corresponding noise reduction effect. Second, the self-attention mechanism in Transformer, which has a better effect in NLP, is used in the model, and the effect is quite good.

The scattering transform in the model can alleviate the problem of timbre similarity to a certain extent, can identify artifacts and has strong robustness to a certain extent. After the scattering transform, 6-layer one-dimensional convolution is used to obtain a larger receptive field, which can reduce the negative impact of invalid time frames while obtaining key information.

At the same time, the model analyzes the self-attention mechanism in the Transformer with the help of the Transformer's success in processing long-term sequences. It was found that it can obtain better global information in the early stage, and can achieve consistency of feature granularity, to achieve a better recognition effect.

To solve the problem of insufficient categories for sound scene recognition. Complements the category in UrbanSound8K with the introduction of ESC-50 and GoogleCommand. This enables the model to fit more classes of sounds and features with different granularizes. This is the ability, robustness and validity of the model to be validated.

REFERENCES

[1] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," in *Proc. IEEE Int. Conf. Multimedia Expo*, New York, NY, USA, Jun. 2009, pp. 1218–1221.

[2] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Acoustic detection of human activities in natural environments," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 686–695, Sep. 2012.

[3] M. Kohiyama, K. Oka, and T. Yamashita, "Detection method of unlearned pattern using support vector machine in damage classification based on deep neural network," *Struct. Control Health Monitor.*, vol. 27, no. 8, p. e2552, Aug. 2020.

[4] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. G. Okuno, "Environmental sound recognition for robot audition using matching-pursuit," in *Proc. Int. Conf. Ind., Eng. Appl. Appl. Intell. Syst. (IEA/AIE)*, Berlin, Germany, 2011, pp. 1–10.

[5] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, May 2011.

[6] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, May 2015.

[7] Y. Zhang, X. Ma, X. Wang, J. Xiang, and W. Wang, "Revisiting the definition of ferroelectric negative capacitance based on Gibbs free energy," in *Proc. 5th IEEE Electron Devices Technol. Manuf. Conf. (EDTM)*, Chengdu, China, Apr. 2021, pp. 1–3.

[8] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, Jun. 2009.

[9] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 477–480.

[10] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, Jul. 2012.

[11] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6445–6449.

[12] S. Abidin, X. Xia, R. Togneri, and F. Sohel, "Local binary pattern with random forest for acoustic scene classification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.

[13] Z. Ren, K. Qian, Y. Wang, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 3, pp. 662–669, May 2018.

[14] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using Gabor filterbank features," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 714–718.

[15] D. Palaz and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," Idiap, Martigny, Switzerland, Tech. Rep. Idiap-RR-23-2015, Jun. 23, 2015.

[16] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 421–425.

[17] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 4624–4628.

[18] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 1021–1028.

[19] P.-G. Noe, T. Parcollet, and M. Morchid, "CGCNN: Complex Gabor convolutional neural network on raw speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7724–7728.

[20] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. 18th Eur. signal Process. Conf.*, Aalborg, Denmark, 2010, pp. 1267–1271.

[21] J. Xu, X. Li, P. Wang, X. Jin, and S. Yao, "Multi-modal noise-robust DDoS attack detection architecture in large-scale networks based on tensor SVD," *IEEE Trans. Netw. Sci. Eng.*, early access, Sep. 12, 2022, doi: 10.1109/TNSE.2022.3205708.

[22] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1543–1551, Sep. 2010.

[23] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018.

[24] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6440–6444.

[25] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[26] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 131–135.

[27] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled AudioSet tagging with attention neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 11, pp. 1791–1802, Nov. 2019.

[28] J. Long, Y. Qin, Z. Yang, Y. Huang, and C. Li, "Discriminative feature learning using a multiscale convolutional capsule network from attitude data for fault diagnosis of industrial robots," *Mech. Syst. Signal Process.*, vol. 182, Jan. 2023, Art. no. 109569.

[29] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.

[30] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[31] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," 2021, *arXiv:2101.08596*.

[32] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 7324–7334.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2016, pp. 770–778.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, New York, NY, USA, Jun. 2017, pp. 4700–4708.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, San Diego, CA, USA, 2017, pp. 5998–6008.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[38] Z. Zhao, Q. Zhang, X. Yu, C. Sun, S. Wang, R. Yan, and X. Chen, "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–28, 2021, doi: 10.1109/TIM.2021.3116309.

[39] Z. He, H. Shao, P. Wang, J. Lin, J. Cheng, and Y. Yang, "Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105313.

[40] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[41] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, QLD, Australia, Oct. 2015, pp. 1015–1018.

[42] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[43] T. Alipour-Fard, M. E. Paoletti, J. M. Haut, H. Arefi, J. Plaza, and A. Plaza, "Multibranch selective kernel networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 1089–1093, Jun. 2021.

[44] J. Lu and S. Steinerberger, "Neural collapse under cross-entropy loss," *Appl. Comput. Harmon. Anal.*, vol. 59, no. 1, pp. 224–241, Jul. 2022.
[45] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 171–175.
[46] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Boston, MA, USA, Sep. 2015, pp. 1–6.
[47] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.
[48] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, London, U.K., Aug. 2017, pp. 1–5.
[49] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019.
[50] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Appl. Sci.*, vol. 8, no. 7, p. 1152, Jul. 2018.
[51] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, 2019.
[52] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Appl. Acoust.*, vol. 167, no. 1, Oct. 2020, Art. no. 107389.
[53] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2019.
[54] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. 39th Int. Conf. Mach. Learn.*, Baltimore, MD, USA, 2022, pp. 17627–17643.
[55] B. Yin, F. Corradi, and S. M. Bohté, "Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks," *Nature Mach. Intell.*, vol. 3, no. 10, pp. 905–913, Oct. 2021.
[56] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2721–2725.
[57] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Guangzhou, China, 2018, pp. 356–367.
[58] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, no. 17, pp. 896–903, Sep. 2021.

SHEN SONG received the B.E. degree in computer science and technology from the Wuhan Institute of Technology, Wuhan, China, in 2018. He is currently pursuing the M.S. degree in computer technology with Wuhan Polytechnic University, Wuhan. His research interest includes artificial intelligence technology and its application.

CONG ZHANG received the bachelor's degree in automation engineering from the Huazhong University of Science and Technology, in 1993, the master's degree in computer application technology from the Wuhan University of Technology, in 1999, and the Ph.D. degree in computer application technology from Wuhan University, in 2010. He is currently a Professor with the School of Electrical and Electronic Engineering, Wuhan Polytechnic University. His research interests include multimedia signal processing, multimedia communication system theory and application, and pattern recognition.

ZHIHUI WEI received the B.E. degree from Wuhan Polytechnic University, Wuhan, China, in 2021, where she is currently pursuing the M.S. degree in computer technology. Her research interest includes artificial intelligence technology and its application.