Applied Intelligence (2024) 54:3017–3039
https://doi.org/10.1007/s10489-023-05138-7

A classification method of marine mammal calls based on two-channel fusion network

Danyang Li[1] · Jie Liao[1] · Hongbo Jiang[1] · Kailin Jiang[2] · Mingwei Chen[1] · Bei Zhou[1] · Haibo Pu[1] · Jun Li[1]

Accepted: 27 October 2023 / Published online: 6 March 2024

Abstract

Marine mammals are an important part of marine ecosystems, and human intervention seriously threatens their living environments. Few studies exist on the marine mammal call recognition task, and the accuracy of current research needs to improve. In this paper, a novel MG-ResFormer two-channel fusion network architecture is proposed, which can extract local features and global timing information from sound signals almost perfectly. Second, in the input stage of the model, we propose an improved acoustic feature energy fingerprint, which is different from the traditional single feature approach. This feature also contains frequency, energy, time sequence and other speech information and has a strong identity. Additionally, to achieve more reliable accuracy in the multiclass call recognition task, we propose a multigranular joint layer to capture the family and genus relationships between classes. In the experimental section, the proposed method is compared with the existing feature extraction methods and recognition methods. In addition, this paper also compares with the latest research, and the proposed method is the most advanced algorithm thus far. Ultimately, our proposed method achieves an accuracy of 99.39% in the marine mammal call recognition task.

Keywords Deep learning · Marine mammal · Sound classification · Two-channel fusion network

1 Introduction

The ocean contains numerous mammals. Marine mammals have always been an important part of marine biodiversity and have a very important role in developing marine ecosystems and maintaining ecological balance. However, with the expansion of human activities, the survival of many marine species is directly or indirectly threatened [1]. To date, the identification and conservation of marine mammals have also remained a hot topic of interest in the field [2]. It is impossible for humans to directly observe the movements of mammals living in the ocean, such as whales and seals, in opaque media, which undoubtedly makes it more difficult to protect marine mammals. However, for most marine mammals, tracking and monitoring tasks through their call signals is one of the most feasible approaches [3]. By implementing effective call classification monitoring tasks for marine mammals, it is important to promote marine mammal conservation and animal welfare.

Most mammals communicate with their own kind through their calls. For example, Seyfarth et al. discovered that animals obtain relevant information for early warning and communication through the same kind of call signals and

Kailin Jiang and Mingwei Chen contributed equally to this work

✉ Jun Li
lijun@sicau.edu.cn

Danyang Li
202005873@stu.sicau.edu.cn

Jie Liao
liaojie@stu.sicau.edu.cn

Hongbo Jiang
202005482@stu.sicau.edu.cn

Kailin Jiang
202003978@stu.sicau.edu.cn

Mingwei Chen
202005560@stu.sicau.edu.cn

Bei Zhou
12801@sicau.edu.cn

Haibo Pu
puhb@sicau.edu.cn

[1] College of Information Engineering, Sichuan Agricultural University, No. 46 Xinkang Road, Ya'an 625000, Sichuan Province, China

[2] College of Science, Sichuan Agricultural University, No. 46 Xinkang Road, Ya'an 625000, Sichuan Province, China

discovered the importance of the same kind of call information in communication [4]. The effect of animal alert messages on foraging behavior was found by Bhattacharjee et al. through simulations using a computer [5]. Takahashi studied information exchange between meerkats and found information exchange among large meerkats [6]. Among marine mammals, call signal transmission has become an important method for recognizing their own kind as well as communicating information. For example, sperm whales communicate with each other by making four different calls [7]. RootGutteridge et al. analyzed the calls of North Atlantic right whales to determine their age by their calls [8]. Tortero-tot et al. used blue whale calls to detect blue whale population distribution in the Southern Indian Ocean [9]. The task of monitoring mammal calls is of great importance for understanding and analyzing their behavior and population status, and scientific and effective detection methods are an important part of supporting this task.

Convolutional neural networks (CNNs) have demonstrated excellent capabilities in many research areas [10]. CNNs can effectively capture the high-dimensional feature information of input data to obtain more expressive information. Its excellent performance was demonstrated in processing animal call classification tasks. Xie et al. achieved success in improving bioacoustic signal classification tasks using CNN structures [11]. Tabak et al. used a CNN model based on the ResNet18 structure to classify 10 bat calls, achieving 92% accuracy [12]. Maegawa et al. proposed a new study for monitoring the presence and reproductive capacity of birds by building CNN models [13]. CNNs have also been widely used in the study of marine mammal calls. For example, Duan et al. proposed using convolutional neural networks for the frame spectrogram data classification task of three marine mammals [14]. Luo et al. accomplished automatic monitoring of toothed whale echolocation clicks by using CNN, which showed excellent stability [15]. Lu et al. proposed a migration learning-based AlexNet approach for marine mammal monitoring tasks, achieving an overall accuracy of 97.42% in three categories [16]. The widespread use of CNNs has proven their suitability for handling animal call classification tasks, as well as for greatly improving animal protection and other welfare causes.

In speech recognition tasks, timing information is extremely important. With the emergence of recurrent neural networks and further research, their application in the speech recognition field is becoming increasingly widespread [17, 18]. For example, Fatih Ertam used a deeper long short-term memory (LSTM) network structure for predicting gender from an audio dataset, and the study successfully predicted gender with a recognition accuracy of 98.4% [19]. Zijiang Zhu et al. proposed a speech emotion recognition model based on Bi-GRU (bidirectional gated recursive unit) and focus loss, which effectively extends short-duration speech samples and uses a focus loss function to address classification difficulties caused by the imbalance of sample emotion categories [20]. Additionally, the attention mechanism achieves excellent performance in capturing dependencies in distant sequences. Compared with the fixed time step of RNN, the attention mechanism can focus on the next prediction task without losing information at longer distances as well. For example, Mohammed M. Nasef et al. proposed two self-attention-based models to provide an end-to-end speech gender recognition system in an unconstrained environment, and their models achieved 95.11% and 96.23% accuracy [21]. Junfeng Zhang et al. extracted speech and text features separately using a two-layer transformer encoder combination model and modeled MoCap using a deep residual shrinkage network with a recognition accuracy of 75.6% [22]. The widespread use of models such as recurrent neural networks and attentional mechanisms that have a good ability to capture temporal information proves their importance for processing speech recognition tasks.

In recent years, marine mammal habitats and populations have been threatened by human activities and other impacts, and an increasing number of researchers are working on marine mammal monitoring and conservation. The main approach of existing studies is the identification and analysis of their calls. However, analyzing all current studies revealed that vocal behavior among mammals is rarely explained in detail, and the description of communication information between mammals is rarely detected [23, 24]. In the existing methods, there are still low model recognition accuracy, few recognition categories, and poor generalization and robustness problems. To address the above problems, this paper proposes a two-way fusion network-based approach for detecting marine mammal classes.

This paper is divided into five sections, the first of which is an introductory section that introduces the work of other researchers and leads us to identify the problem. The second section presents the theoretical and applied possibilities of our approach. The third section shows the experiments and results of our method and provides a rational analysis. The fourth section discusses our work. The fifth section is the conclusion.

1.1 Main innovations and contributions

The main innovations and contributions of this paper are as follows:

1. In the previous audio recognition tasks, most of them take a single network structure for recognition. Such as CNN, RNN, transformer, etc. But the single network structure cannot extract the information in the audio thoroughly, such as CNN cannot use the timing information in the

sound signal effectively. In this paper, a two-way parallel fusion network structure is constructed, and by combining the improved MG-Resnet and MG-Transformer structures, the fusion network can have the ability to capture audio high-dimensional features and utilize audio timing information at the same time.

2. In this paper, we construct a multi-granularity joint layer for aiding multi-classification tasks. We construct a coarse-grained layer and a fine-grained layer for the species data to be identified by using the "kingdom, phylum, class, order, family, genus and species" division. The coarse-grained layer corresponds to the "family" of the species, and the fine-grained layer corresponds to the "genus" of the species, and the coarse-grained a priori judgments are used to consolidate the decisions of the fine-grained layer in subsequent work. Our proposed multigranularity fusion layer has strong generalizability and can be applied to other studies.
3. In this paper, we design a new audio feature called energy fingerprint. This feature contains a large amount of information about marine mammal calls, such as energy, frequency, timing information, etc., which identifies the vocal expression ability of marine mammals. After experiments, the feature is proved to have good performance.

2 Related work

Convolutional neural networks use feature invariance to overcome the inherent diversity of speech signals in the audio recognition domain and can learn data-driven, highly representative hierarchical audio features from sufficient training data and have shown excellent performance in the audio domain. However, it still suffers from problems such as nontrivial feature selection, environmental noise or degradation of accuracy due to intensive local computation. In recent years, many studies have been devoted to solving these problems. Cao et al. conducted an appropriate combination of CNNs and hand-designed features, using the minimum redundancy and maximum correlation (mRMR) algorithm as a criterion for selecting the best set of hand-designed features, and achieved higher recognition accuracy than using CNNs alone [25]. Xu et al. designed and implemented a CNN-based acoustic classification system. Additionally, to improve the accuracy in noisy environments, a multiview CNN framework is proposed that contains three convolution operations and three different filter lengths to extract short-, medium- and long-term information simultaneously. The architecture achieves better accuracy and significantly outperforms traditional CNN classification models when ambient noise dominates the audio signal (low SNR) [26]. Shawn et al. investigated the size of the training set and label vocabulary and found that state-of-the-art image networks were able to achieve superior results in audio classification compared to simple fully connected networks or earlier image classification architectures. Some performance improvements were derived when training on larger training and label sets, and regularization reduced the gap between models trained on smaller datasets and 70 M datasets [27]. Nanni et al. studied a collection of classifiers for automatic animal audio classification using different data enhancement techniques to train convolutional neural networks (CNNs) and showed that training different CNN animal audio classification models worked better than standalone classifiers [28]. Xie et al. proposed a new feature set by first applying a sliding window to the audio waveform to obtain the plus-window signal, where the five windows with the highest energy are selected. An orthogonal matching trace is applied to these windowed signals to extract the significant Gabor atoms. A multiwindow scale frequency map is constructed as an input feature for three different CNNs, and experiments on two classification datasets also demonstrate the effectiveness of their framework in complementing traditional audio time-frequency representations [29].

After the transformer was proposed in the natural language processing field, it was successively introduced to the speech recognition and computer vision fields and has performed very well. Since then, transformers have been increasingly used in speech recognition. pan et al. extracted latent representations by sampling a subset of patches with low attention weights in the transformer encoder model. and using environmental information for fusion with tokens with high attention weights to improve the distinguishability of dynamic attention fusion models [30]. Gong et al. used self-attention-based neural networks in the audio domain, where the audio spectrogram transformer (AST) achieved excellent performance in various audio classification tasks [31]. Lee et al. proposed a dual cross-modal (DCM) attention scheme that exploits both audio context vectors from video queries and video context vectors using audio queries and introduces a connectionist temporal classification (CTC) loss to the attention-based model to enforce the required monotonic alignment in AVSR [32]. Wang et al. proposed a distributed visual channel coding scheme based on a multimodal converter and deep joint source channel coding-based distributed audiovisual parsing network (DAVPNet), which is used to enhance attentional computation between audiovisual events [33].

In recent years, good progress has been made in other audio recognition tasks. For example, Dufourq et al. evaluated the ability of state-of-the-art migration learning models to classify animal calls in four bioacoustic datasets and the impact of various modeling decisions on recognition accuracy, fully developing migration learning in a PAM-based environment while simplifying the CNN design architecture [34]. Oikarinen et al. introduced an end-to-end feedforward convolutional neural network Oikarinen et al. introduced

an end-to-end feedforward convolutional neural network that can reliably classify the source and type of animal calls in noisy environments after training with imperfectly labeled datasets, providing an idea for researchers interested in studying vocalizations [35]. Salamon et al. compared a 'shallow learning' approach based on unsupervised dictionary learning with a deep convolutional neural network enhanced with data to improve the ability to monitor biodiversity [36].

3 Our method

We propose a pioneering parallel recognition model. It consists of two phases. The first stage extracts energy fingerprint features from the original audio. The energy fingerprint is a new feature we propose for audio signals, which aims to retain as much information as possible from multiple perspectives in the audio signal, such as frequency, energy, and timing information. The good performance of this feature is verified in the experimental section. The energy fingerprint features are fed into ResNet18 with a combined coarse and fine granularity layer added for training (MG-ResNet) to finally obtain the classification vector in the first stage. The second stage extracts the traditional audio signal features from the original audio: MFCC. The MFCC features will be fed into the transformer (MG-Transformer) with the coarse-fine granularity joint layer for training, and finally, the classification vector of the second stage is obtained. We design a trainable fusion layer to receive the first- and second-stage classification vectors and fuse them. After that, the final classification results are output. Figure 1 illustrates the general architecture of our approach.

3.1 Feature extraction

In training deep learning models, the input features must be insensitive to phase, so the task of signal processing cannot be accomplished using raw audio. Raw audio data are often high-dimensional and contain considerable redundant information caused by strong correlations, so direct training is often inefficient. Therefore, feature extraction of the raw audio is essential.

Audio feature extraction can streamline the sampled signal of the original waveform, thus accelerating the machine's understanding of the semantic meaning in the audio. To obtain the audio features that work best, we extracted nine features that are mainstream in audio data: chromatographic information, constant-Q chromatographic information, normalized chromatographic information, Meier spectral information, MFCC, spectral contrast, tonal center of mass, local autocorrelation of the onset intensity envelope, and Fourier tachogram. These nine features are explained in Table 1.

Figure 2 shows a sample from our data, which was first preemphasized in the initial processing to increase the energy of the high-frequency part of the signal. Given a time-domain input signal x[n], the signal after preemphasis is:

$$y[n] = x[n] - \alpha x[n-1], 0.9 \leq \alpha \leq 1.0 \tag{1}$$

The windowing process is carried out for the preemphasized sound signal, and to facilitate the subsequent extraction of various features, we make the value of the signal at the window boundary approximate to 0 so that the signal tends to be a periodic signal, and the windowing function is as follows:

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right), & 0 \leq n \leq L-1 \\ 0, \quad \text{otherwize} \end{cases} \tag{2}$$

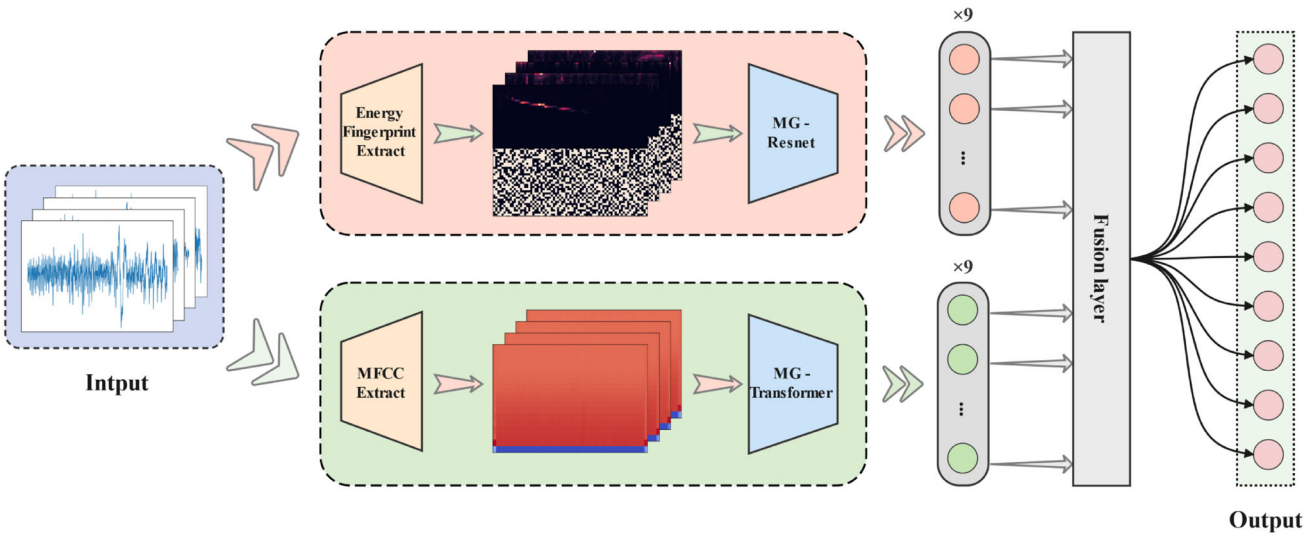

Fig. 1 Methodology Overview

Table 1 Feature Introduction

Name	Description	Related work
MFCC	MFCC are cepstrum parameters extracted in the frequency domain of the Mel scale, which describes the nonlinear properties of human ear hearing.	Hasan M R et al. [37]
Chromagram	Chromagram, also called a pitch class profile, shows the energy distribution of a sound over a series of pitches. This series of pitches is usually the twelve pitches of western music.	Ellis et al. [38]
Spectral contrast	The spectral contrast divides each frame of the spectrum into subbands. For each subband, the energy contrast is estimated by comparing the average energy of the top part bits with the average energy of the bottom part bits.	Jiang et al. [39]
Mel scaled spectrogram	The sound spectrum map of the raw data is often a very large map, which is often transformed into a mel spectrum by passing it through mel-scale filter banks in order to obtain sound features of the right size.	Hasan M R et al. [40]
Constant-Q chromagram	The chromatogram after constant Q-transformation. The constant Q-transform transforms time series to the frequency domain and is widely used in the processing of sound data.	Manzo-Martinez A et al. [41]
Tonal central features	In audio theory, tonal central features project chromatic features onto a six-dimensional basis, representing perfect fifths, minor thirds and major thirds as two-dimensional coordinates, respectively.	Harte et al. [42]
Tempogram	Tempogram is obtained by calculating the local autocorrelation of the starting intensity envelope.	Grosche et al. [43]
Chroma energy normalized	Chroma energy normalized features are robust in terms of dynamics, timbre and intelligibility. Therefore, they are commonly used in audio matching and retrieval applications.	Meinard Müller and Sebastian Ewert . [44]
Fourier tempogram	The fourier tempogram is obtained by performing a short-time fourier transform on the starting intensity envelope.	Grosche et al. [45]

To visualize these nine features, we extracted features from one preprocessed signal in the dataset and visualized the feature matrix. Figure 3 below shows the effect of feature visualization, along with an explanation of the meaning of each mapping.

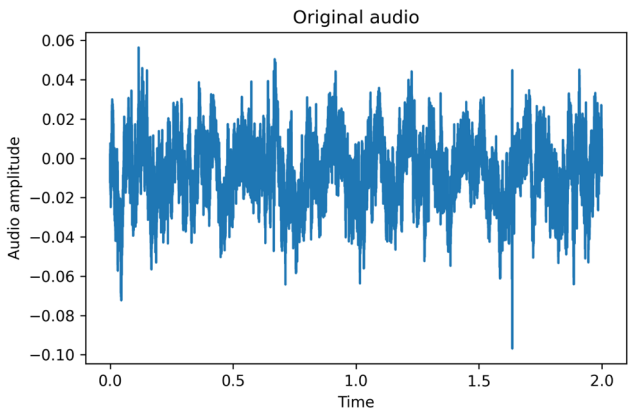

Fig. 2 Original audio samples

Mel-frequency spectrogram is obtained by taking logarithmic transform after Fourier transform and Mayer filter. It is usually based on time and Mayer frequency as axes, and the color depth represents the energy intensity of the frequency. MFCC is based on the Mel spectrum graph and is obtained by discrete cosine transform, using time and Mel frequency as coordinate parameters. In fact, it compresses the low-frequency information of the Mel spectrum graph to extract the main features. Spectral contrast enhances the feature discrimination of the spectrogram and the high-frequency characteristics of the signal. It is obtained by calculating the STFT representation and calculating spectral energy, spectral logarithmic energy, and spectral contrast. Tempogram extracts the initial intensity envelope, performs preprocessing to remove DC offsets, and then calculates the autocorrelation between samples near a certain time point in the signal from local autocorrelation. In the graph, the

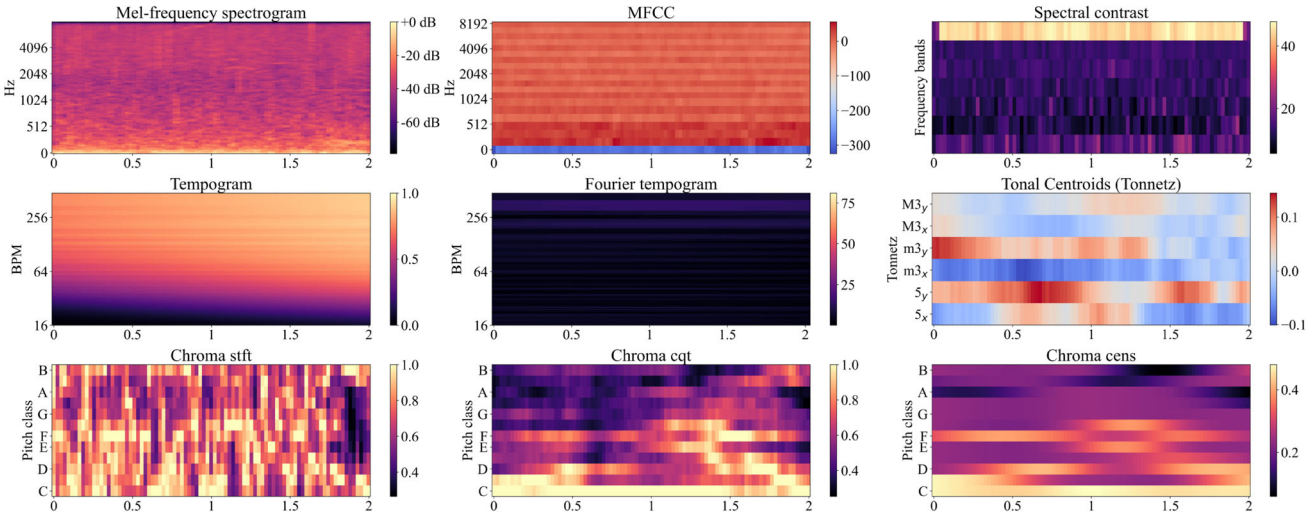

Fig. 3 From left to right, in order from top to bottom, they are Mel-frequency spectrogram, MFCC, Spectral contrast, Tempogram, Fourier tempogram, Tonal central features, Chromagram, Constant-Q chromagram and Chroma energy normalized

horizontal axis represents time and the vertical axis represents local autocorrelation coefficients. Fourier tempogram performs time-frequency analysis on audio signals, typically performing frame division processing on audio signals, then calculating their short-time Fourier transform, and drawing images with time as the horizontal axis, frequency as the vertical axis, and energy intensity as the color. Tonal Centroids calculates the pitch spectrum weighted center frequency of each frame of data to obtain the corresponding tone centroid, and plots it in chronological order. Chroma sift is a visual representation of audio signals in terms of time and frequency. It represents the volume levels at different frequencies through color graphics, presenting the spectral characteristics of audio signals. Constant cqt uses the CQT spectrum of the calculated audio to achieve an exponential distribution of the center frequency, connecting the amplitude spectra of all frames together to form a color representation of the amplitude of each frequency along the time and frequency axes. Chroma cens is usually obtained by normalizing the spectrum obtained by performing a short time Fourier transform on audio. Each pixel represents the energy value within a frequency and time period.

3.2 Energy fingerprint

Traditional audio features such as mel-scaled spectrograms, MFCCs, and diagrams are often related to only a single piece of information in the sound pattern. Therefore, we hope to construct a new type of acoustic signature that contains frequency, energy, and timing to enhance the weaknesses that exist in a single signal. We also want this feature to reflect the uniqueness of different species' vocalizations, which we call the energy fingerprint, and the construction process is shown in Fig. 4.

First, we want to minimize the impact of too long and too short audio on the final classification. Therefore, we use the atomic frame streaming strategy in the construction of the energy fingerprint, i.e., the original audio is divided into atomic frames of the same size, after which a series of changes are made to the atomic frames to obtain the atomic features, and the final features are obtained by combining these atomic features. Common atomic features may not contain enough information to support the model to identify them. However, under normal circumstances, the audio to be recognized is composed of hundreds and thousands of atomic frames, containing enough of them for efficient and reliable recognition [46].

Next, we performed a Fourier transform on the data of the atomic frames to calculate their spectral information. The Fourier transform is calculated as shown in (3) and (4).

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \tag{3}$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{j\omega t} d\omega \tag{4}$$

An arbitrary signal is obtained as a sequence of imaginary numbers after Fourier transform. The amplitude of the spectrum is obtained by modeling each imaginary number obtained. The calculation formula is shown in (5) and (6).

$$F\left(e^{j\omega}\right) = a + ib \tag{5}$$

$$|F(e^{j\omega})| = \sqrt{a^2 + b^2} \tag{6}$$

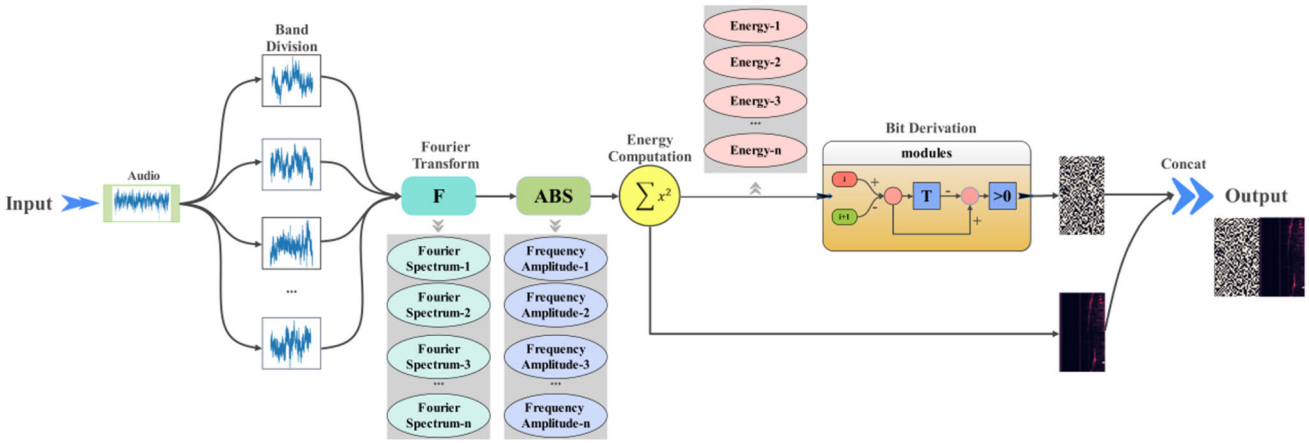

Fig. 4 Construction flow of energy fingerprint

To express the energy information more precisely, we divided the atomic spectrum into 65 spectral bands again, and for each band, the energy blocks were calculated. According to Perseval's theorem, the total energy of the signal can be obtained either as the integral of the energy per unit of time over the whole time or as the integral of the energy per unit of frequency over the whole frequency range. Therefore, the energy value of each energy block can be calculated by (7).

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} |F(e^{j\omega})|^2 d\omega \tag{7}$$

Thus, after combining 65 energy blocks, we obtain the energy information of the atomic spectrum. This energy information of the atomic spectrum will form an energy matrix, as shown in Fig. 5. In this way, the frequencies and energies are fully extracted while retaining some timing information, which will be an important feature of our energy fingerprint.

As we mentioned above, we wanted to obtain a feature that was unique to the vocalizations of the species. It is easy to understand that much of the uniqueness of species vocalizations is reflected in the temporal information of the sounds. However, unlike human vocalizations, it is difficult to obtain the semantic meaning of marine mammal sounds. For the language spoken by humans, it is extremely easy to obtain semantic information. Therefore, we need to find a property that will change frequently during the vocalizations of marine mammals while at the same time reflecting a certain temporal sequence. This property is energy information. Marine mammals express different semantic meanings during vocalization, and the energy changes are obvious. In turn, the energy change is highly distinguishable among different species, such as the period, change interval, and energy mean. This is one of our reasons for constructing energy fingerprints.

The energy matrix presented above consists of 100 × 65 energy blocks arranged and combined according to the time dimension and frequency domain dimension, and each energy block is obtained by its corresponding spectral band calculation. To extract as much information as possible about its timing, we have to consider not only the relationship between neighboring energy values in the same column of energy blocks (each energy value represents energy in a different frequency domain). We also have to consider the relationship between energy blocks with different times and the same frequency domain. Therefore, we choose two differential calculations in different dimensions to capture the connection between each energy block and its neighboring energy blocks. A matrix containing only 0 and 1 is finally obtained, as shown in Fig. 6. The calculation formula is shown in (8), where x is the coordinate in the time dimension

Fig. 5 Energy matrix

Fig. 6 Zero-One matrix

```
Input: original audio of shape (L,), the number of divided sub-audio n, audio length m and energy's length m'
Output: energy fingerprint of shape
function EnergyFingerprintGenerater
    SplitList←BandDivision(Input,n,m)
    //The length of every audio in SplitList with the length of n is m.
    for i←1 to n do
        SplitList[i]←FourierTransform(SplitList[i])
        SplitList[i]←ABS(SplitList[i])
    //Get Energy
    Energy←EnergyComputation(SqlitList,m')
    //Bit Dervation begins
    Fingerprint←Pad(SplitList,0,1)   //Pad(SplitList,0,1) means to fill a circle of 0 around SplitList.
    for i←1 to n do
        for j←1 to m do
            judge←Fingerprint[i][j]- Fingerprint[i][j+1]- Fingerprint[i-1][j]+Fingerprint[i-1][j+1]
            if judge>0 then Fingerprint[i][j]←1
            else Fingerprint[i][j] ←0
    //Concatenate
    output←concat([Fingerprint,Energy], 2)    // concat([A,B], 1) means to concatenate two matrix along the second dimension.
return output
```

Fig. 7 The pseudo-code for extracting audio energy fingerprint features

of the energy matrix and y is the coordinate in the frequency domain dimension of the energy matrix.

$$F(x,y)=\begin{cases}1, & [E(x,y)-E(x,y+1)]-[E(x-1,y)-E(x-1,y+1)]>0\\ 0, & [E(x,y)-E(x,y+1)]-[E(x-1,y)-E(x-1,y+1)]\le 0\end{cases} \tag{8}$$

This kind of binary matrix contains the vocal information of the creature. Eventually, we splice the energy matrix with the zero-one matrix to obtain the final energy fingerprint.

The following Fig. 7 shows the pseudo-code for extracting audio energy fingerprint features for reference.

3.3 Multigranularity joint layer

In traditional classification tasks, the classified objects are often in a hierarchical relationship, which leads to a neural network with approximate learning ability for each class. In real life, these taxonomic objects often have corresponding coarse-grained categories. For example, marine mammals are very diverse, and biologists have developed a "kingdom, phylum, class, order, family, genus and species" classification to better distinguish them.

Many effective models and ideas in deep learning are mostly derived from bionic ideas. For example, the convolution operation to extract image features mimics the process of receiving a picture in primates [47]. Attentional mechanisms mimic attention in human vision [48]. The combination of coarse-grained and fine-grained categorization also mimics the process of perceiving categories of a thing, from abstract to figurative. The difference in coarse-grained classification object features is generally large. For example, it is easy to distinguish between cars and people. The differences in fine-grained classification object features are smaller. Thus, for things that have both coarse-grained categories and fine-grained categories, the neural network can perceive both coarse-grained attributes and fine-grained attributes. Such an approach is clearly more reasonable than direct classification [49].

YOLO9000 can detect more than 9000 object classifications in real time thanks to its designed WordTree structure [50]. WordTree is used to mix the data among the detection dataset and recognition dataset. Additionally, YOLO9000 can obtain different levels of class information of a sample by WordTree. Ming Sun et al. proposed a multiattention multiclass constraint structure [51]. Their model learns the multiattention feature regions of each input image by the one-squeeze multiexcitation (OSME) model and uses the multiattention multiclass constraint (MAMC) structure to guide the extracted attention features to correspond to the category labels. Therefore, we borrowed these ideas of using both coarse and fine granularity and added a "multigranularity joint layer" to the neural network. The structure is shown in Fig. 8 below.

In this structure, we modify the final fully connected layer of the neural network so that it maps a tensor with lengths of the number of fine-grained categories and the number of coarse-grained categories, which are the inputs to our fine-grained and coarse-grained layers, respectively. It is worth noting that the data we use here still have only one label, the fine-grained label, and the capture of coarse-grained information relies on the coarse-grained layer for implementation.

The coarse-grained layer uses a softmax operation on the input data to obtain the probability that a sample belongs to a coarse-grained category. The fine-grained layer groups the input data, divides the fine-grained data belonging to the same coarse-grained category into groups, and later performs a softmax operation on each group. Finally, the category probability information is obtained by multiplying the coarse-grained probabilities with the corresponding fine-grained probabilities, and the following equation represents

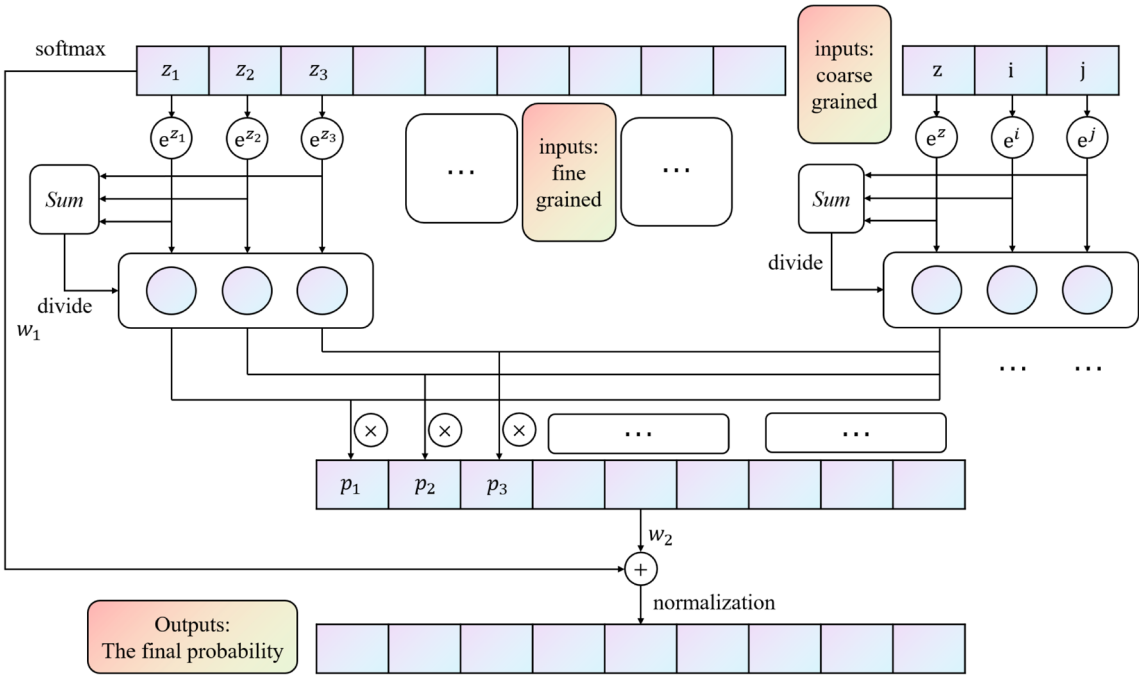

Fig. 8 Schematic of multi-grain size joint layer structure

the formula for calculating the category probability of killer whales:

$$p_1(\text{ Killer whale }) = p(\text{ Killer whale } \mid \text{ Whale }) \times p(\text{ Whale }) \quad (9)$$

In addition, we need to consider a special case where the coarse-grained layer may output incorrect coarse-grained category probabilities, although the probability of this case is extremely low. Therefore, we design a kind of residual structure that is used to perform the softmax operation directly on the input of the fine-grained layer, assign the weight ε and then operate with the matrix where p_1 is located, i.e.,

$$p_2(\text{ Killer whale }) = p_1(\text{Killer whale }) + \varepsilon \times p_0(\text{ Killer whale }) \quad (10)$$

$p_1(Killerwhale)$ is the probability of the killer whale calculated by softmax directly for the input of the fine-grained layer.

Finally, we normalize the matrix where p_2 is located to obtain the final probability matrix of the sample.

$$p_3(\text{ Killer whale }) = \frac{p_2(\text{ Killer whale })}{\sum p_2} \quad (11)$$

During the actual training process, the neural network will gradually notice the correctness of the learned features. The more correct the coarse-grained category division to which the fine-grained category belongs, the more likely the fine-grained category will be assigned correctly, and although we only used fine-grained labels, the final classification result is still very satisfactory.

3.4 MG-Resnet module for MG-ResFormer

Most traditional sound recognition methods rely on some feature extraction recognition operators, such as hidden Markov, but the model machines constructed by these methods depend on the data representation they are designed for; it is time-consuming and difficult to design a suitable operator for complex and variable tasks and they do not generalize [52]. Additionally, due to the influence of the data acquisition environment and equipment, the performance of the model will be severely restricted; that is, the operator designed manually is not robust.

Sound recognition by a CNN has good robustness and it can gradually extract features with more advanced semantic information abstractly [53]. For example, its first layer may notice the overall sound amplitude, the second layer can capture the information of different species when the sound transitions and the higher layers can capture richer and more abstract high-dimensional features, which are tracked and captured by the neural network as the depth of the network continues to increase to obtain the final classification results.

Additionally, convolutional neural networks have translation invariance. In the field of image classification, this property means that targets in an image can be success-

fully recognized whether they are panned, rotated, or scaled, or even in different lighting conditions and viewing angles. In audio classification, this feature has a further advantage: when the quantity of data is large enough, any piece of data input to the CNN can be considered to have been generated by the previous data translation transformation, but the system produces the same response. This effect is what we want to obtain.

However, there is also a disappearing gradient risk. when applying CNN directly to audio recognition, so we consider using a residual neural network in this experiment. The CNNs greatest role is to effectively solve the problems of depth degradation, and exploding and disappearing gradients that tend to occur during network forward propagation. The more layers the CNN has, the more abstract features the network can extract, the richer the semantic information, and the more likely it is to cause disappearing and exploding gradients. The solution to exploding and disappearing gradients is usually to use regularization, but this leads to degradation, that is, the accuracy of the model on the test set saturates or even decreases. The deep degradation phenomenon is reflected in the fact that when we add new convolutional layers to the CNN, it is likely that the newly added layers will not learn deeper features but only replicate shallow features. In ResNet, the residual module is a good solution to the deepening network and degrading model problems. The residuals are changed from absolute to relative quantities by introducing direct connections, and the calculation becomes much easier. The constant mapping is equivalent to a gradient high-speed channel that allows the neural network to connect in alternate layers, weakening the strong connection between each layer and avoiding the disappearing gradient problem to achieve a network with deeper layers.

Equation 12 is an expression for the residual structure, where x and y are the input and output of the module, respectively, and F is the convolution operation. w_i is a trainable parameter, such as a convolution kernel. The rightmost x of the equation is the residual connection. The x in parentheses is the input of the residual module, and $F(x_i, w_i)$ denotes the output of the backbone after module processing.

$$y_i = F(x_i, w_i) + x_i \tag{12}$$

Shape mismatch may occur when performing tensor addition. Therefore, it is necessary to add training parameters w_s to the residual connection to adjust the size of the residual connection.

$$y_i = F(x_i, w_i) + w_s x_i \tag{13}$$

To facilitate the derivation to illustrate the advantage of the residual module, we ignore w_s and obtain (14).

$$x_{i+1} = x_i + F(x_i, w_i) \tag{14}$$

This indicates that no parameter adjustment is needed for the residual connection. Then, we can obtain recursive (15).

$$\begin{aligned} x_{i+2} &= x_{i+1} + F(x_{i+1}, w_{i+1}) \\ &= x_i + F(x_i, w_i) + F(x_{i+1}, w_{i+1}) \end{aligned} \tag{15}$$

The recursive formula is generalized to obtain (16).

$$x_n = x_l + \sum_{i=l}^{n-1} F(x_i, w_i) \tag{16}$$

Backpropagating Equation xxx-1 is calculated to obtain (17).

$$\frac{\partial \text{ loss}}{\partial x_l} = \frac{\partial \text{ loss}}{\partial x_n}\frac{\partial x_n}{x_l} = \frac{\partial \text{ loss}}{\partial x_n}\left(1 + \frac{\partial}{\partial x_n}\sum_{i=l}^{n-1} F(x_i, w_i)\right) \tag{17}$$

Here, we find that $\frac{\partial}{\partial x_n}\sum_{i=l}^{n-1} F(x_i, w_i)$ is not always -1 in a batch, which means that the gradient will not be 0 when the weights are small enough, i.e., the gradient will not disappear. Combining the generality of the above derivation and the characteristics of the audio data, this residual structure advantage can still be retained for audio recognition tasks, so we consider applying it to our task.

We conducted experiments comparing multiple features with multiple convolutional neural networks, and we analyzed the experimental data and found that the innovative multigranularity joint layer of ResNet18 has excellent performance with energy fingerprinting, so this combination will be the CNN module in our fusion network.

ResNet18 is a neural network with the number of weight layers set to 18, which was proposed by Kaiming He's team in 2015 [54]. ResNet18 uses the BasicBlock residual module, which contains a residual branch and a shortcut branch, allowing the network to be trained very deeply because it has an additional shortcut branch for passing low-level information compared to traditional convolutional structures. Resnet18 uses the BasicBlock residual module, which contains a residual branch and a short-cut branch, allowing the network to be trained very deeply because it has an additional short-cut branch for passing low-level information than traditional convolutional structures.

MG-ResNet refers to the structure of ResNet18 and adds a multigrain joint layer, whose structure and parameters of each layer are shown in Fig. 9, and has a total of five convolution modules. First, 7×7 convolution is performed on the input, then four residual convolution modules containing two build-blocks are performed, followed by two parallel fully connected layers, and the two outputs are fed into the multigrain joint layer to finally obtain the probability prediction of the model on the samples.

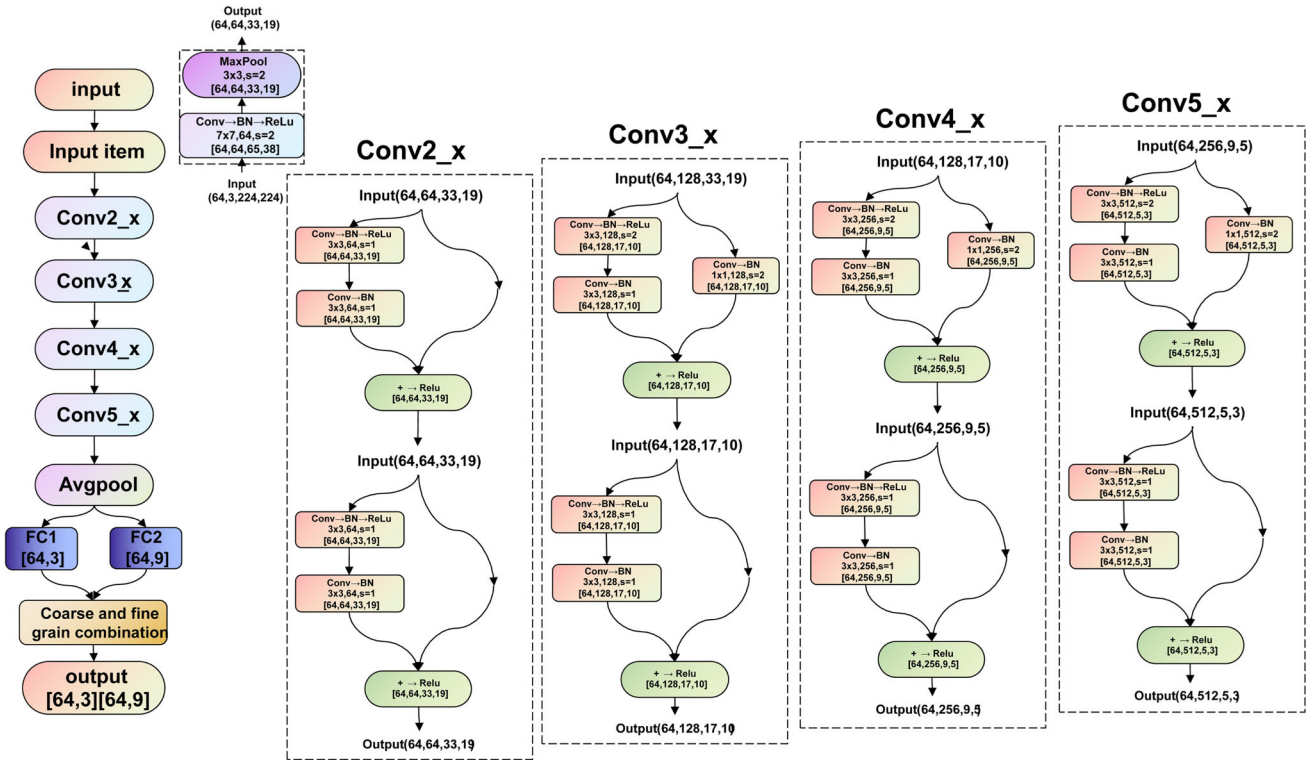

Fig. 9 MG-Resnet structure

3.5 MG-Transformer module for MG-ResFormer

Speech signals are a common time series, and pronunciation signals have extremely strong dependencies on each other. In early classification tasks, recognition methods such as RNN are widely used, but they still have this drawback. Although RNNs account for the dependencies existing between adjacent time series, they often underperform for long-distance temporal information. In the call signal of marine mammals, each sample often contains much information. With a transformer, multifocus is good at capturing dependencies in distant sequences, allowing us to focus on the next prediction task without losing information over longer distances [55]. This is very suitable for the task of classifying marine mammal calls. In this module, we use a different model structure than for tasks such as machine translation and eliminate the decoding layer. This is because we do not need the model to be verbal in this task, so we only use the encoding layer.

In the encoding layer, the multihead attention will slice the input data into multiple parts and use them for the input of each attention head. Each attention head will process the input using the attention formula, which is shown in (18).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

Q is the query matrix, K is the content that we want to focus on, and QK^T is the matrix dot product. Softmax is the normalization operation because we end up normalizing it to between 0 and 1, using decimals between 0 and 1 to reflect the importance of each part. V is our input value, that is, the input value to complete the attention distribution operation. It is worth noting that Q, K and V are not the original values but are obtained by the corresponding matrix multiplication transformations. The matrix multiplication transformations used are W_q, W_k, and W_v. The purpose of this is to complete the input vector mapping using the matrix dot product. This mapping unifies the different Q, K, and V into a uniform type that is convenient for computation. After that, the transformer aggregates the output results of each multihead attention head and uses the parameters to perform dimensioning, as shown in (19) and (20). Therefore, the multihead attention mechanism can effectively increase the parallelism of the transformer by improving the computational speed.

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \tag{19}$$

$$\text{where} \quad \text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{20}$$

We could have used an RNN to learn sound sequences, but the RNN can only learn to predict frequency changes based on adjacent time steps. In contrast, the transformer's powerful parallel computing with multihead attention allows the network to look at multiple previous time steps when

predicting the next step, solving the problem of RNNs. For speech classification tasks, the transformer's powerful global sensing capability is its advantage over RNNs in addressing similar tasks.

We construct a two-layer transformer encoder structure for sensing global features using MFCC (64 × 75) as the input to the model, and the main structure of this encoder is shown in Table 2 below.

As shown in Table 2, we first pool the input features to reduce the sensitivity of the features to different locations and then input them to the encoding layer. In the encoding layer, we use a multihead attention mechanism to extract different feature signals, and by dividing the features, the model enhances the attention to global features, which is useful. In the subsequent processing, we add two linear layers to extract coarse-grained and fine-grained probabilities, such that the model maintains efficient recognition performance for the basic classification task. By using this module, the model can better perceive global features and achieves better performance for multicategory tasks.

3.6 MG-ResFormer

In this paper, a new and powerful parallel fusion network is proposed, which consists of MG-ResNet and MG-Transformer. This architecture takes full advantage of MG-ResNet's high-dimensional feature extraction and MG-Transformer's global time-series information capture capability. In contrast to previous sound recognition classification tasks, where a single network always has the disadvantages of inadequate feature extraction or failure to notice temporal information, the fusion architecture proposed in this paper almost perfectly extracts the valid information in an audio segment. The structure diagram of the MG-ResFormer proposed in this paper is given in Fig. 10.

Table 2 Transformer Encoder Structure

Operator	Output Shape
MaxPool2d	[64, 1, 64, 37]
Multi-head-Attention	[-1, 2, 64]
Dropout	[64, 2, 64]
LayerNorm	[64, 2, 64]
Linear	[64, 2, 64]
Dropout	[64, 2, 64]
Linear	[64, 2, 64]
Dropout	[64, 2, 64]
LayerNorm	[64, 2, 64]
Multi-head-Attention	[-1, 2, 64]
Dropout	[64, 2, 64]
LayerNorm	[64, 2, 64]
Linear	[64, 2, 64]
Dropout	[64, 2, 64]
Linear	[64, 2, 64]
Dropout	[64, 2, 64]
LayerNorm	[64, 2, 64]
Linear	[64, 9]
Linear	[64, 3]

The energy fingerprint feature of the audio is input into MG-ResNet, and its loss function is denoted as $loss^1$. In MG-Transformer, the MFCC features of the audio are input, and its loss function is denoted as $loss^2$. In the fusion module, the probabilities of MG-ResNet and MG-Transformer outputs are used as inputs to fit with the label information after one-hot encoding, and the loss is denoted as $loss^3$. The final loss function is denoted as (21).

$$\text{Loss} = a \times \text{loss}^1 + b \times \text{loss}^2 + c \times \text{loss}^3 \tag{21}$$

We set the hyperparameters a, b, and c as their weights before $loss^1$, $loss^2$, and $loss^3$, respectively. This is because there may be different fitting speeds for different modules. Different fitting speeds of each module will lead to a poor fusion effect, i.e., the fusion layer will focus too much attention on one network method and ignore the other method. To ensure that each module fits at a similar speed to achieve the best overall model, we use a grid search method to search for combinations of the above hyperparameters.

To reduce the large number of computations caused by grid search, we set $a, b, c \in (10^k, k = 0, 1, 2, 3, 4)$. We plotted Fig. 11 to visualize the data from the grid search. The values on the axes in the figure represent the values of the hyperparameters searched, and the color of the coordinate points represents the accuracy achieved by the model for that hyperparameter combination. With the grid search, we found the highest accuracy that the model can achieve to be 99.39%.

As shown in Fig. 10 above, the fingerprint features and MFCC features of an audio sample are input into MG-ResNet and MG-Transformer, respectively, and after forward propagation of the two-way network, two matrices for predicting the sample category probability are output and the two matrices are frozen at this time. Because these two matrices carry the computation process from the two-way, loss3 will change the parameters of the two-way network when it is backpropagated, while $loss^1$ and $loss^2$ have already completed the update of the two-way network; then, the backpropagation of $loss^3$ will have a negative impact on the two-way network.

Since the gradient is backpropagated according to the loss, the abovementioned loss calculation process is correct. After careful bias calculation, $loss^1$, $loss^2$ and $loss^3$ contained in the loss are responsible for backpropagating the module to which they belong.

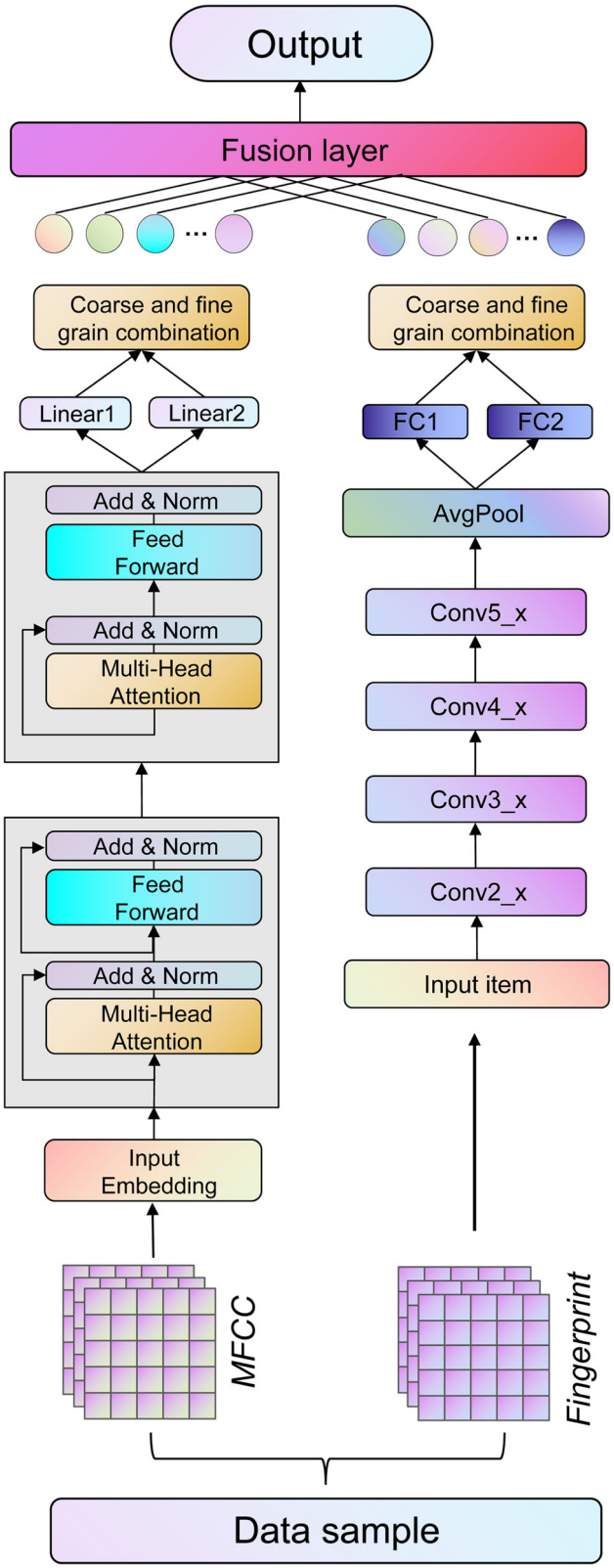

Fig. 10 MG-ResFormer structure

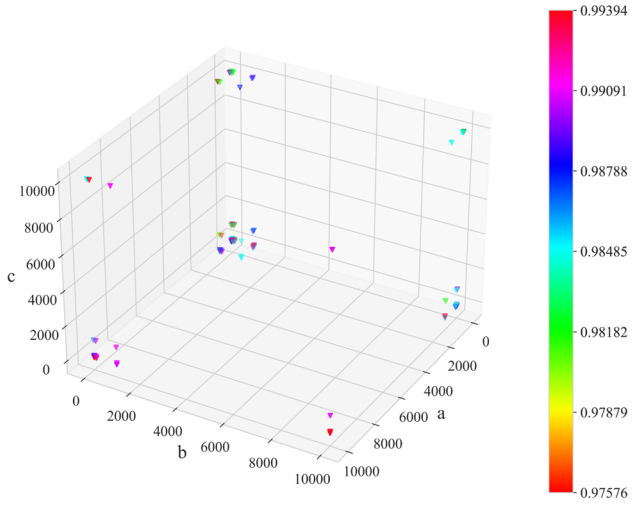

Fig. 11 Grid search visualization

In the fusion module shown in Fig. 12, we set up nine neurons, and the probability values of the nine categories of the two network outputs, denoted as Input1 and Input2, pass through nine neurons and nine pseudoneurons, respectively, where Input1 performs the Hadamard product operation with the neurons and Input2 performs the Hadamard product operation with the pseudoneurons, after which the two sets of probability values obtained are summed for the normalization operation to obtain Output1. Here, the pseudoneurons are explained: there are also 9 pseudoneurons, and the value of each pseudoneuron depends on the value of the neuron at the corresponding position. As shown in (22), β_{Ti} denotes neurons involved in training and β_{Fi} denotes pseudoneurons that are not involved in training.

$$\beta_{Fi} = 1 - \beta_{Ti} \tag{22}$$

To ensure that our fusion structure has the correct impact, we also use the same idea of class residuals, and the final probability of the output is normalized by adding Input1 and Input2. After calculating (23), the final output is obtained.

$$\text{Output} = \frac{\text{Output 1} + \text{Input 1} + \text{Input 2}}{\sum(\text{Output 1} + \text{Input 1} + \text{Input 2})} \tag{23}$$

Both normalization operations performed in the fusion layer are L^2 parametric normalization. We define the L^2 parametrization of the vector $x(x1, x2, ..., xn)$ to be normalized as $norm(x) = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$.

We need to normalize x to the unit L^2 parametrization, i.e., create a mapping from x to x' such that the L^2 parametriza-

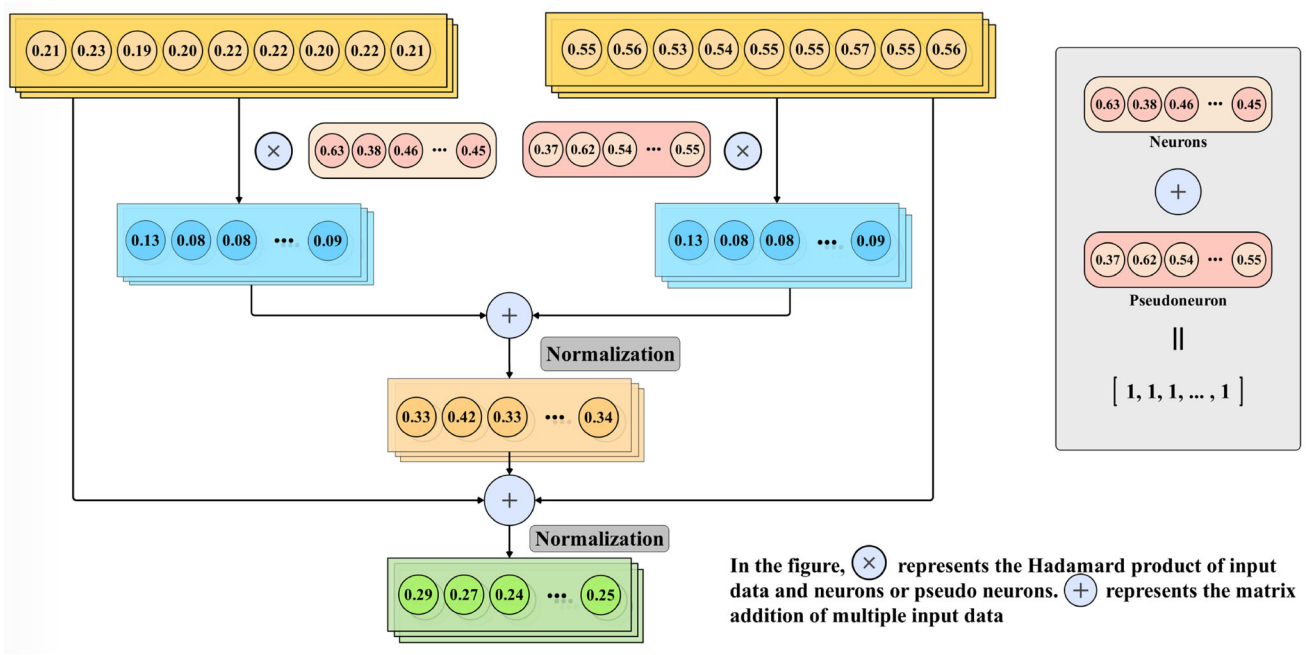

Fig. 12 Fusion layer structure

tion of x' is 1. Thus, the derivation in (24) is obtained.

$$\begin{aligned} 1 = \text{norm}\left(x'\right) &= \frac{\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}}{\text{norm}(x)} \\ &= \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{\text{norm}(x)^2}} \\ &= \sqrt{\left(\frac{x_1}{\text{norm}(x)}\right)^2 + \left(\frac{x_2}{\text{norm}(x)}\right)^2 + \ldots + \left(\frac{x_n}{\text{norm}(x)}\right)^2} \\ &= \sqrt{x_1'^2 + x_2'^2 + \ldots + x_n'^2} \end{aligned} \tag{24}$$

The final normalized (25) is obtained.

$$x_i' = \frac{x_i}{\text{norm}(x)} \tag{25}$$

The normalization process is necessary to prevent the singular samples from affecting the model fit. L^2 parametric normalization has the characteristic of nonsparsity and can take advantage of more features in the matrix instead of simply ignoring them.

4 Experiments and results

4.1 Datasets

We used information on nine classes of marine mammals publicly available in the Watkins Marine Mammal Sound Database, which provides marine mammal sound recordings from 1940 to 2000 and offers three options for conducting experiments: best clips, all cuts and masters, with all clips containing approximately 15,000 sound clips and masters containing nearly 1,600 complete tapes [56]. We chose the best cut section as the data used in this paper, and because the data chosen are from different equipment in different geographical locations in different generations, our work is highly generalizable. Additionally, to facilitate the generation of energy fingerprints with the same network input data scale, we split the varying audio lengths evenly into 2 s of audio. The training set and test set are divided according to the ratio of 4:1, and the specific statistical information is shown in Fig. 13.

4.2 Evaluation indicators

To better evaluate the prediction quality and generalization ability of the model in this paper, we chose four evaluation metrics: ACC, AUC, mAP, and $f1_{score}$. Before calculating these metrics, several concepts need to be clarified. TP denotes the positive samples predicted by the model to be in the positive category. TN denotes the negative samples predicted by the model to be in the negative category. FP denotes the negative samples predicted by the model to be in the positive category. FN denotes the positive samples predicted by the model to be in the negative category.

ACC, the most commonly used classification evaluation metric, is primarily used to indicate the number of samples

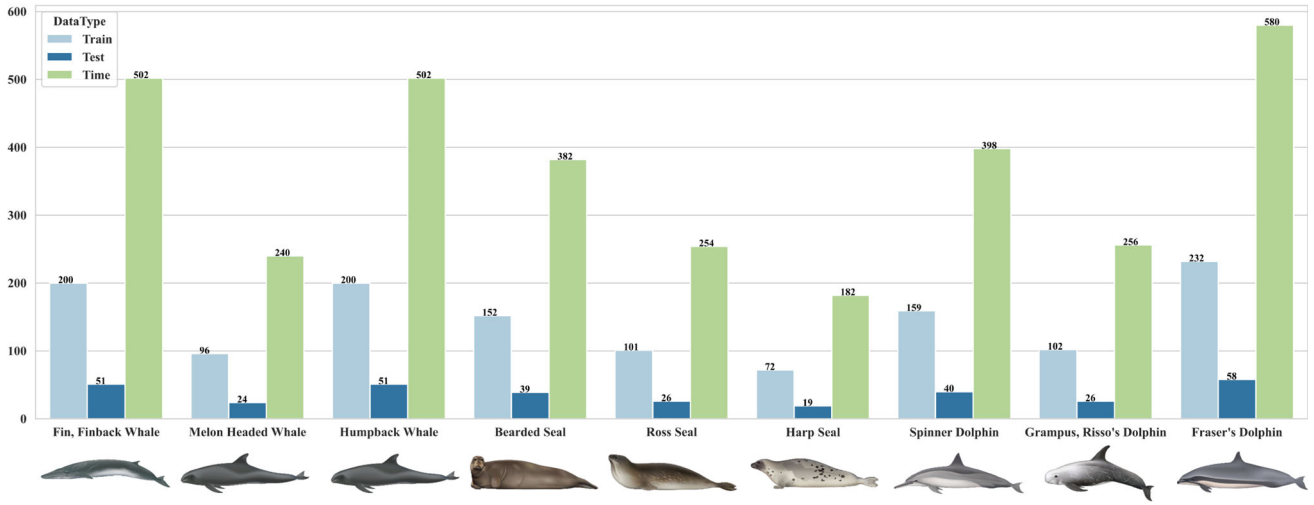

Fig. 13 Data presentation

correctly predicted as a percentage of the total sample size. the formula is as follows:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)} \tag{26}$$

AUC, area under the curve, represents the sum of the areas under the ROC curve, a metric that quantifies the performance of the classification algorithm. In contrast, the ROC curve consists of two indicators: TPR, the true positive rate, and FPR, the false positive rate. Given a classifier, multiple sets (FPR,TPR) can be obtained by changing the probability threshold for the classifier to determine positive and negative samples. By connecting these points, the ROC curve required to calculate the AUC is obtained. The specific formula is as follows:

$$TPR = \frac{TP}{TP + FN} \tag{27}$$

$$FPR = \frac{FP}{FP + TN} \tag{28}$$

$$AUC = \int_0^{+\infty} FPRdTPR \tag{29}$$

The mAP is an indicator for multiple categories; it requires AP information for each category. The AP is calculated using precision and recall. Precision is the proportion of samples where both predicted and actual values are true to those predicted as true. Recall is the proportion of samples where both predicted and actual values are true to the overall true sample.

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{30}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{31}$$

$$AP = \int_0^1 P(R)dR = \sum_{k=0}^{n} P(k)\Delta R(k) \tag{32}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{33}$$

The $F1_{score}$ accounts for both accuracy and recall and is defined as the summed average of accuracy and recall. It is widely used in model evaluation.

$$f1_{\text{score}} = \frac{2 \times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{34}$$

4.3 Experimental settings

All networks in this paper are written using the mainstream deep learning framework PyTorch. The hardware and software configurations of the devices are shown in Table 3.

The relevant parameters of the deep learning neural network model in this paper are shown in Table 4.

Due to the limitations of the experimental environment of the device, the batch size and the number of iterations of the

Table 3 Hardware and software configuration table

Configuration	Detail
CPU	AMD Ryzen 5 2600X Six-Core Processor
RAM	64G
Graphics Card	NVIDIA GeForce RTX 3090
Operating System	64-bit Windows 10
CUDA	11.6
Programming Language and Version	Python 3.8

Table 4 Network model parameter list

Parameters	Value
Batch Size	32
Epochs	50
Ratio of Training Set to Test Set	4:1
Initial Learning Rate	0.001

deep learning model in this paper are set to 32 and 50, respectively. Additionally, Adam is chosen for model optimization because of its high computational efficiency, faster convergence and good interpretability of the hyperparameters. The initial learning rate is set to 0.001. The training and test sets are divided in a ratio of 4:1 for all datasets used in this paper.

4.4 Experiment

A total of four experiments were designed to demonstrate the effectiveness of the proposed method. The first experiment is conducted by combining the network capturing temporal features and the convolutional neural network with each feature separately. In this way, the combination of network and features that captures temporal permitting best and the combination of convolutional neural network that extracts high-dimensional features best can be selected respectively. The two combinations obtained from this experiment will be used as the basis for the subsequent validation experiments of the two-way fusion network. In the second experiment, we test the accuracy improvement brought by adding a multi-granularity joint layer model based on the combination obtained in the first experiment and prove its effectiveness. In the third experiment we verify that the fusion network has a significant improvement in accuracy compared to the single-way network. In the fourth experiment we validate the generalization of the method by applying the proposed method to multiple sound datasets.

1. **Network and feature combination selection experiments**

   In Experiment 1, we conducted experiments on selecting two-way networks with feature combinations. The experiment was designed to obtain the optimal choice of combinations for capturing local features versus capturing timing information. Among the neural networks capturing timing, we chose four networks, RNN, GRU, LSTM, and transformer, for the experiments [19, 55, 57]. Among the convolutional neural networks, we chose AlexNet, VGG11, ResNet18, and GoogLeNet for the experiments [58–60].

   We used accuracy as an evaluation metric. In the comparison of multiple sets of experiments, we obtained the two combinations with the highest accuracy: ResNet with the energy fingerprint feature and transformer with the MFCC feature. The experimental data are shown in Figs. 14 and 15.

The result shows that ResNet has a very powerful residual structure and is simpler than other network structures, making it easier to achieve better results in this task. Additionally, the energy fingerprint generation process extracts the main information and features of the original audio and presents them in a lower dimension, which facilitates further feature extraction and learning by the convolutional neural network. The transformer's unique network structure and its powerful multiheaded attention mechanism can capture the timing information in the audio features, and the MFCC features generated after inverse spectrum analysis of the Meier spectrum highlight the discrete time-domain information of the convolutional signal, and the combination of the two gives excellent results.

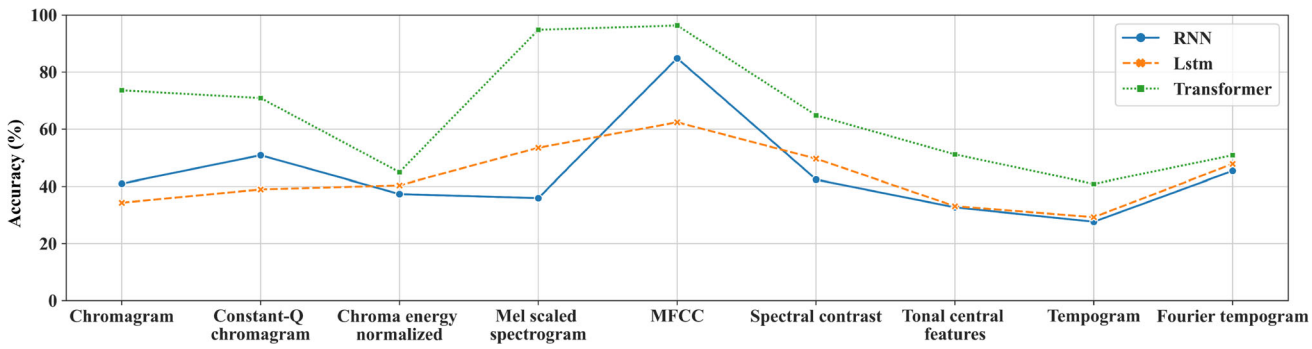

Fig. 14 Comparison experiment of RNN, LSTM and Transformer

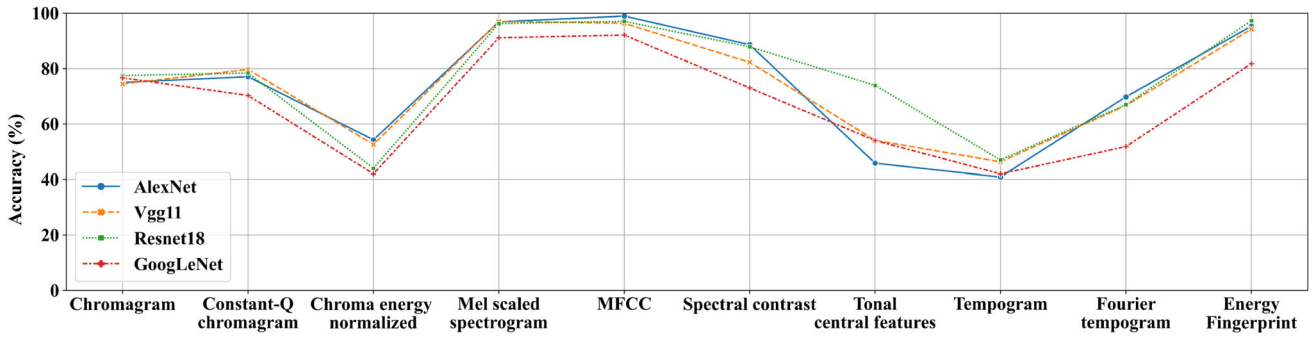

Fig. 15 Comparison experiment of AlexNet, Vgg11, Resnet18 and GoogLeNet

2. **Comparison of models with the addition of multi-granularity joint layers**

   In Table 5, we show the performance of ResNet18 compared to that of MG-ResNet and that of the transformer compared to that of MG-Transformer. The multigranularity union layer achieved significant performance in tasks with multiple classifications and with hierarchical connections between classes.

3. **Comparison of fusion network and general Network**

   In Experiment 3, we tested the MG-ResFormer proposed in this paper, and we can see in Table 6 that it performed extremely well in the audio nine-classification task for marine mammals. In this task, its ACC, AUC, Map, and $f1_{score}$ reached 99.39, 99.99, 99.92 and 99.28, respectively. Compared with classification networks commonly used in the same field, this network achieved significant improvement.

We plotted the accuracy images of MG-ResFormer during the training process. Figure 16 shows that MG-ResNet and MG-Transformer have oscillations in test set accuracy during training. In contrast, MG-ResFormer has a smooth increase in test set accuracy. This is the subtlety of the fusion layer. Both MG-ResNet and MG-Transformer have the oscillation problem, which indicates that there may be some odd samples in the dataset, causing the two-way network to not fit smoothly during the training process. We considered the problem of odd samples at the beginning of the design, as they are commonly found in sound recognition datasets. Therefore, we designed the fusion layer containing two normalizations. The classification vectors output from the two networks goes through two normalization operations during the fusion process. The normalization speeds up the gradient descent to find the optimal solution and speeds up the training network convergence. This is one of the important reasons why the MG-ResFormer test set accuracy rises more smoothly. Additionally, the neurons in the fusion layer perform a weighting calculation on the classification vectors of the two network outputs. The MG-ResFormer accuracy is almost always higher than that of MG-ResNet and MG-Transformer, which proves that the fusion of the two networks plays a complementary role in improving the MG-ResFormer accuracy.

The final model can achieve high accuracy mainly for the following reasons.

We designed the backbone network from two different perspectives (CNN and transformer) initially, hoping that these two different classifiers can produce complementary effects and capture more effective features from the input data.

Convolutional neural networks, when learning traditional audio features, lose a large amount of timing information. Therefore, we designed a new feature for audio signals: the energy fingerprint. This timing information is designed in the energy fingerprint by differential computation, which can be

Table 5 The effect of multi-granularity joint layers on accurcy

Network/Indicators	ACC(%)	AUC(%)	mAP(%)	$f1_{score}$(%)
Resnet18	96.67	99.91	99.26	99.26
MG-Resnet	97.27	99.93	99.35	96.54
Transformer	95.45	99.87	98.90	94.11
MG-Transformer	96.36	99.85	98.63	95.76

Table 6 Converged networks versus general networks

Network/Indicators	ACC(%)	AUC(%)	mAP(%)	$f1_{score}$(%)
MG-ResFormer	99.39	99.99	99.92	99.28
MG-Resnet	97.27	99.93	99.35	96.54
MG-Transformer	96.36	99.85	98.63	95.76
InceptionV3	94.62	98.97	99.03	94.43
EfficientNet	95.39	99.14	99.02	95.78

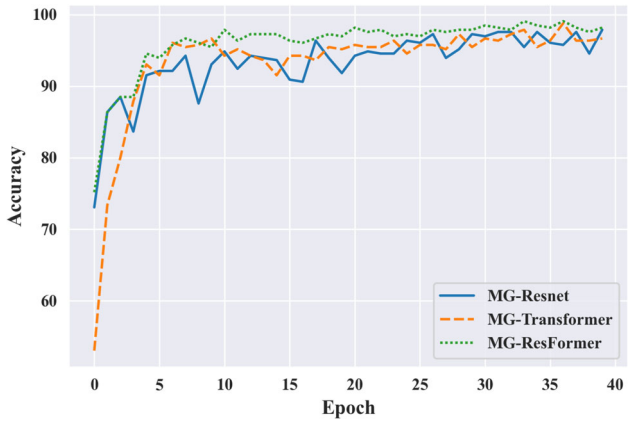

Fig. 16 Changes in test set accuracy during training for MG-ResFormer, MG-Resnet and MG-Transformer

effectively perceived by the convolutional neural network. Therefore, this feature can retain as much timing information as possible during the gradual increase in dimensionality after the input to the convolutional neural network.

We conducted many experiments to find the best combination, i.e., ResNet with energy fingerprint features and transformer with MFCC features obtained the best results. We designed the fusion layer. After the two parallel networks output their classification vectors separately, the fusion layer back calculates the weighting of the classification vectors.

We designed the fusion layer. After the two parallel networks output their classification vectors separately, the fusion layer back calculates the weighting of the classification vectors. This weight is obtained by continuously training the neurons in the fusion layer. The computation also goes through two normalization operations to accelerate the gradient descent. In this way, the confidence level of the two-way network on different categories can be obtained, resulting in more accurate results.

4. Generalizability experiments

To demonstrate the excellent generalization of our work to similar tasks, we validated it on other audio datasets. The audio datasets used are the bird sounds dataset (from the bird audio data publicly available at xeno-canto) [61] and the urban sounds dataset (UrbanSound8K) [62].

Xeno-canto is a website dedicated to sharing bird sounds from around the world and contains the sounds of 10,357 species. We selected nine of these species for our validation experiments: pale rainbows, brown rock nectar suckers, wrens, European silkies, green-footed sandpipers, western ruffed grouse, green whistling thrushes, gray-breasted hydrangeas, and brown whistlers. The length of each audio file varies from a dozen seconds to a few minutes and was cut to facilitate the generation of our subsequent features, resulting in 11,670 features for training purposes. The energy fingerprint size was 130×75, and the MFCC size was 64×75.

Urbansound8K is a widely used public dataset for automated urban environmental sound classification studies. The dataset contains a total of 8,732 annotated sound segments ($<= 4s$) in 10 categories: air conditioning, car sirens, children playing, barking dogs, boreholes, idling engines, gunshots, handheld drills, police sirens and street music. The final dataset was segmented to also generate energy fingerprints of size 130×75 and MFCC of size 64×75 for experiments. We divide the urban sounds dataset and the bird sounds dataset into training and testing sets according to 4:1, and the specific statistical information is shown in Figs. 17 and 18.

We conducted experiments on the proposed method using the urban sounds dataset and bird calls dataset in this paper. Similarly, we used the ACC, AUC, mAP, and $f1_score$ as evaluation metrics to evaluate the performance of the model.

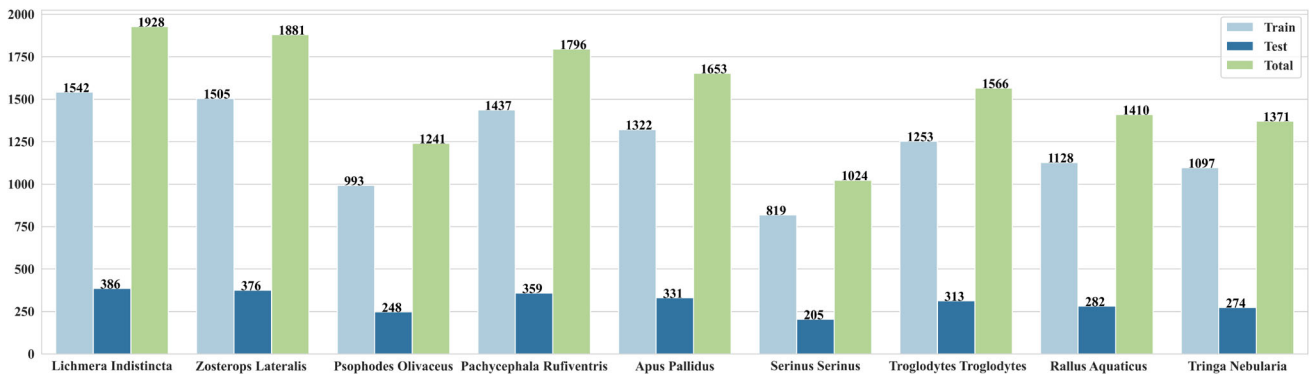

Fig. 17 Bird Call Statistics

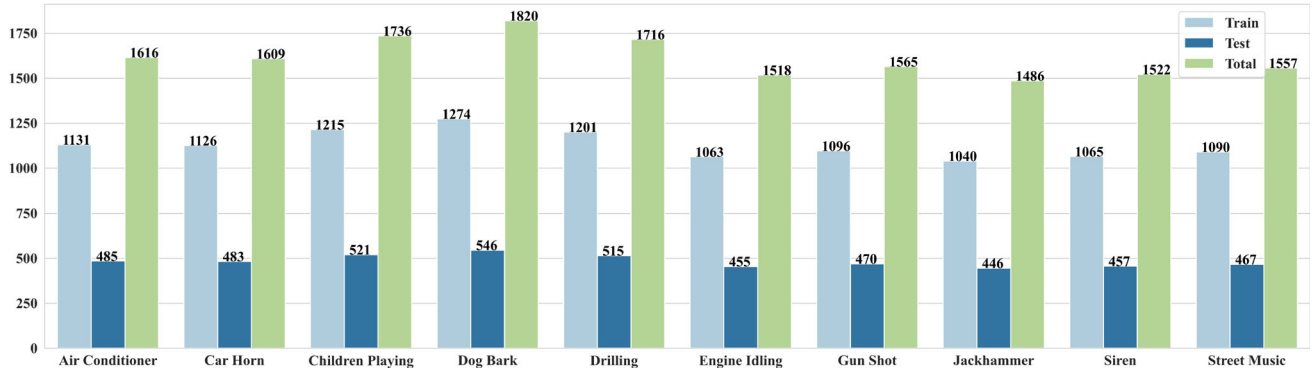

Fig. 18 City Voice Statistics

We obtained the experimental results as shown in Table 7. Experiments using different datasets show that our method can effectively extract high-dimensional information from different speech signals, and significantly improve speech classification task accuracy through multistrength joint layers, network fusion and multiple features, which are also highly applicable to different data.

5 Discussion

In this paper, we propose a new acoustic feature, the energy fingerprint, and apply it to the task of acoustic recognition of marine mammals. The energy fingerprint was constructed with the intention of including as much information as possible about the original sound: frequency, energy, and the potential information about the unique vocal pattern of the organism, so that it is well represented as an input feature for each marine mammal. However, this has the disadvantage that cutting atomic frames destroys some of the timing information of the original audio, but this is perfectly compensated for by the fusion network structure in this paper, which will be explained in more detail later. Additionally, humans perceive categories in terms of granularity, so neural networks can also learn granularity information. To make better use of the relationship between classes, i.e., the hierarchical relationship between kingdoms, phyla, families, genera and species", this paper proposes a multigranularity joint approach to classification. In addition, this paper constructs a powerful MG-ResFormer fusion network structure, which consists of MG-ResNet and MG-Transformer. Since this paper addresses a classification task and does not require the model to be linguistically expressive, the MG-Transfomer consists of the encoder structure in the transformer.

The MG-ResFormer takes two sanple features, the energy fingerprint and the Meier cepstral coefficients, and sends the energy fingerprint to the MG-ResNet, whose local feature extraction capability is sensitive to the category information contained in the energy fingerprint, which is not available in the transformer. The Merle cepstrum coefficients are fed into the MG-Transformer. We designed this one-way network to capture the timing information contained in the features and to compensate for the lack of sound timing extraction in MG-ResNet. The reason for not using RNN networks for this task is that RNN networks can only predict at a fixed time step; however, most of the sounds made by marine mammals are high-frequency information, and at a fixed time step, almost all of this high-frequency information would be wasted, as demonstrated in the experiments. The transformer's multi-headed attention mechanism, on the other hand, can focus on global temporal information, which means that information from any two time steps can be related in some way. Finally, a network fusion layer is designed to fuse the category decisions of two networks for one sample. The output of MG-ResNet and MG-Transformer are used as input to flow through the neurons and pseudoneurons of the fusion layer, respectively, and after a series of operations, the final category probability values are output. This fusion takes advantage of the complementary nature of the CNN and the transformer to perfectly classify the marine mammal sounds.

Table 7 Generalizability Verification Experiments

Generalizability exprriments/Indicators	ACC	AUC	mAP	$f1_{score}$
Birds 4 Classification	97.16	99.63	99.02	96.28
City 4 Classification	92.16	98.95	96.84	91.77
Birds 9 Classification	95.63	99.32	97.42	95.51
City 10 Classification	83.47	97.83	91.97	83.12

Table 8 Our work in comparsion with the work of other research scholars

Study	Number of Classes	Method	Accuracy
Tao Lu et al. [16]	3 classes	Spectrogram, Alexnet	97.42%
Fernando RubénGonzález-Hernández et al. [63]	11 classes	Octave analysis, Parallel neuural networks	90.00%
Dexin Duan et al. [14]	3 classes	Spectrogram, CNN	91.28%
Marek B.Trawicki. [64]	9 classes	MFCC, HMM	82.72%
Our Work	9 classes	Our Method	99.39%

We compared the method proposed in this paper with the studies of other scholars in the same field, as shown in Table 8, which contains data on the research team, the number of research categories, the method used, and the performance. Our proposed method has better performance on the marine mammal call recognition task. Not only did we identify more marine mammal categories than most of the works, but we also achieved a higher accuracy. For the work performed by Fernando RubénGonzález-Hernández et al., we identified only two fewer species, but with 9.39% higher accuracy. In the future, we will improve our method to further expand the number of recognition categories and guarantee accuracy.

Although the original intention of this paper was to use a state-of-the-art approach to identify and classify marine mammal sounds, the method is equally applicable to a number of other sound recognition tasks. To test this idea, this paper experimented with the proposed method using data from Urbansound8K, a widely used public dataset for automatic urban environmental sound classification studies, and xeno-canto, a world bird sound sharing site. The accuracy of the model proposed in this paper reached 93.29% and 83.02% in the urban sound10 classification task and the bird call9 classification task, respectively. This proves that the method can be widely used for sound classification tasks in various fields and is highly advanced. We also found that the classification results of urban sound data are not as satisfactory as those of other types of data. We attempted to analyze the phenomenon and determined the reasons for this result. Urban sound data contain sounds such as car whistles, idling engines, and children playing. These sounds contain a large amount of noise, which seriously interferes with the neural network in extracting the effective features in the sounds, and we believe this is the main reason for the poor results.

Although the method proposed in this paper achieved objective results, it still has certain drawbacks. First, in the combined coarse and fine granularity layer, we only rely on the neural network to perceive the coarse and fine granularity of the data and do not impose an explicit penalty mechanism on the neural network in the perception process. An in-depth study from this perspective may be able to further improve the model's performance. Second, the overall architecture of the neural network in this paper is a two-channel parallel architecture, which may have unexpected effects if some layers or some parameters in the two-channel network are allowed to be shared.

We encourage other researchers to propose more novel approaches based on this paper. The method proposed in this paper and its excellent call recognition performance is of great significance and opens new directions for marine mammal conservation. Additionally, the method bridges the gap between visual methods in marine mammal detection tasks. Marine mammals mostly live seat depths of 200-1,000 meters underwater. Due to various reasons, such as light, impurities, and tides, it is difficult for camera equipment to capture images that can accomplish the marine mammal visual monitoring task.

The method proposed in this paper is groundbreaking for the marine mammal call identification task. In this paper, we innovatively propose the concept of energy fingerprints and coarse and fine intensity joint layers and combine the improved ResNet and transformer two-channel parallel networks through a fusion layer. Our future work will continue to explore the application of multichannel deep fusion networks for sound recognition.

6 Conclusion

In this paper, we first propose a new audio feature for the marine mammal call classification task, energy fingerprinting, which contains a large amount of information about the original audio, such as energy, frequency, timing information and vocal characteristics specific to different species. The energy fingerprint contains more information than the current mainstream audio features. Moreover, marine mammal classification includes a "kingdom, phylum, family, genus, and species" connection between classes, which is often overlooked by existing research methods, so we designed a multigranularity joint layer to guide the neural network to learn this potential connection. Additionally, to compensate for the shortcomings of convolutional neural networks that cannot effectively utilize temporal information, we designed a two-way fusion network structure, MG-ResFormer, which extracts two different dimensional features of the original

audio (energy fingerprint and mel inversion coefficient) by feeding them into different network structures to extract different information (MG-ResNet and MG-Transformer) and finally, performing the classification task by fusing the high-dimensional features. This approach makes almost perfect use of the information present in the audio. We performed two normalization operations on the input vectors in the fusion layer to make the MG-ResFormer fit better and effectively avoid the oscillation problem during the training process. Finally, MG-ResFormer achieves 99.39% accuracy in the classification task of nine marine mammal calls. In addition, MG-ResFormer can be widely applied to other audio classification tasks. We tested our proposed model on the city sound dataset and the bird calls dataset and showed through experiments that MG-ResFormer still achieves excellent generalization ability for different data and is strongly portable with little influence from the data. The work in this paper is a pioneering exploration in the marine mammal call recognition and sound recognition field. This work can make a great contribution to the monitoring and identification of marine mammals while providing new ideas for sound recognition tasks.

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Danyang Li, Jie Liao and Hongbo Jiang. The first draft of the manuscript was written by Danyang Li, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Subsidy for the University Student Innovation Training Program (No.202210626003).

Availability of data and materials All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval The project uses publicly available data that has been subject to an exemption.

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or nonfinancial interest in the subject matter or materials discussed in this manuscript.

References

1. Avila IC, Kaschner K, Dormann CF (2018) Current global risks to marine mammals: Taking stock of the threats. Biol Conserv 221:44–58, 01 May 2018
2. Brando S, Broom DM, Acasuso-Rivero C, Clark F (2018) Optimal marine mammal welfare under human care: Current efforts and future directions. Behav Process 156:16–36, 01 Nov 2018
3. Verfuss UK et al (2018) Comparing methods suitable for monitoring marine mammals in low visibility conditions during seismic surveys. Mar Pollut Bull 126:1–18, 01 Jan 2018
4. Seyfarth RM, Cheney DL, Bergman T, Fischer J, Zuberbühler K, Hammerschmidt K (2010) The central importance of information in studies of animal communication. Anim Behav 80(1):3–8, 01 July 2010
5. Bhattacharjee S, MacPherson B, Wang RF, Gras R (2019) Animal communication of fear and safety related to foraging behavior and fitness: An individual-based modeling approach. Ecol Inform 54:101011, 01 Nov 2019
6. Takahashi DY (2018) Animal Communication: Chit-Chat in Meerkats. Curr Biol 28(22):R1298–R1300, 19 Nov 2018
7. Jiang J et al (2021) Study of the relationship between sound signals and behaviors of a sperm whale during the hunting process. Appl Acoust 174:107745, 01 Mar 2021
8. Root-Gutteridge H, Cusano DA, Shiu Y, Nowacek DP, Van Parijs SM, Parks SE (2018) A lifetime of changing calls: North Atlantic right whales, Eubalaena glacialis, refine call production as they age. Anim Behav 137:21–34, 01 Mar 2018
9. Torterotot M, Samaran F, Stafford KM, Royer J-Y (2020) Distribution of blue whale populations in the Southern Indian Ocean based on a decade of acoustic monitoring. Deep Sea Res Part II: Top Stud Oceanogr 179:104874, 01 Sept 2020
10. Alzubaidi L et al (2021) Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J big Data 8(1):1–74
11. Xie J, Hu K, Guo Y, Zhu Q, Yu J (2021) On loss functions and CNNs for improved bioacoustic signal classification. Ecol Inform 64:101331, 01 Sept 2021
12. Tabak MA, Murray KL, Reed AM, Lombardi JA, Bay KJ (2022) Automated classification of bat echolocation call recordings with artificial intelligence. Ecol Inform 68:101526, 01 May 2022
13. Maegawa Y et al (2021) A new survey method using convolutional neural networks for automatic classification of bird calls. Ecol Inform 61:101164, 01 Mar 2021
14. Duan D et al (2022) Real-time identification of marine mammal calls based on convolutional neural networks. Appl Acoust 192:108755, 01 April 2022
15. Luo W, Yang W, Zhang Y (2019) Convolutional neural network for detecting odontocete echolocation clicks. The J Acoust Soc Am 145(1):EL7–EL12
16. Lu T, Han B, Yu F (2021) Detection and classification of marine mammal sounds using AlexNet with transfer learning. Ecol Inform 62:101277, 01 May 2021
17. Toderici G et al (2017) Recurrent Neural Network Regularization
18. Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Comput 9(8):1735–1780
19. Ertam F (2019) An effective gender recognition approach using voice data via deeper LSTM networks. Appl Acoust 156:351–358, 15 Dec 2019

20. Zhu Z, Dai W, Hu Y, Li J (2020) Speech emotion recognition model based on Bi-GRU and Focal Loss. Pattern Recog Lett 140:358–365, 01 Dec 2020
21. Nasef MM, Sauber AM, Nabil MM (2021) Voice gender recognition under unconstrained environments using self-attention. Appl Acoust 175:107823, 01 April 2021
22. Zhang J, Xing L, Tan Z, Wang H, Wang K (2022) Multi-head attention fusion networks for multi-modal speech emotion recognition. Comput & Ind Eng 168:108078, 01 June 2022
23. Flack JC (2013) Animal communication: hidden complexity. Curr Biol 23(21):R967–R969
24. Pika et al (2018) Taking turns: bridging the gap between human and animal communication. Proceedings of the Royal Society. Biol Sci
25. Cao Z, Principe JC, Ouyang B et al (2015) Marine animal classification using combined CNN and hand-designed image features[C]. OCEANS 2015-MTS/IEEE Washington. IEEE, pp 1–6
26. Xu W, Zhang X, Yao L et al (2020) A multi-view CNN-based acoustic classification system for automatic animal species identification[J]. Ad Hoc Netw 102:102115
27. Hershey S, Chaudhuri S, Ellis DP W et al (2017) CNN architectures for large-scale audio classification[C]. 2017 IEEE International conference on acoustics, speech and signal processing (icassp). IEEE, pp 131–135
28. Nanni L, Maguolo G, Paci M (2020) Data augmentation approaches for improving animal audio classification[J]. Eco Inform 57:101084
29. Xie J, Zhu M (2022) Sliding-window based scale-frequency map for bird sound classification using 2D-and 3D-CNN[J]. Expert Syst Appl 207:118054
30. Pan H, Xie L, Wang Z (2022) Plant and Animal Species Recognition Based on Dynamic Vision Transformer Architecture[J]. Remote Sensing 14(20):5242
31. Gong Y, Lai CI, Chung YA et al (2022) Ssast: Self-supervised audio spectrogram transformer[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10):10699–10709
32. Lee YH, Jang DW, Kim JB et al (2020) Audio-visual speech recognition based on dual cross-modality attentions with the transformer model[J]. Appl Sci 10(20):7263
33. Wang P, Li J, Ma M et al (2022) Distributed Audio-Visual Parsing Based On Multimodal Transformer and Deep Joint Source Channel Coding[C]. ICASSP 2022-2022 IEEE International conference on acoustics speech and signal processing (ICASSP). IEEE, 4623–4627
34. Dufourq E, Batist C, Foquet R et al (2022) Passive acoustic monitoring of animal populations with transfer learning[J]. Eco Inform 70:101688
35. Oikarinen T, Srinivasan K, Meisner O et al (2019) Deep convolutional network for animal sound classification and source attribution using dual audio recordings[J]. The J Acoust Soc Am 145(2):654–662
36. Salamon J, Bello JP, Farnsworth A et al (2017) Fusing shallow and deep learning for bioacoustic bird species classification[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 141–145
37. Hasan MR, Jamil M, Rahman M (2004) Speaker identification using mel frequency cepstral coefficients[J]. variations, 1(4):565–568
38. Ellis D (2007) Chroma feature analysis and synthesis[J]. Resources of laboratory for the recognition and organization of speech and Audio-LabROSA 5
39. Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, Lian-Hong Cai (2002) Music type classification by spectral contrast feature. In Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International conference on. IEEE, 1:113–116
40. Hasan M R, Jamil M, Rahman M. Speaker identification using mel frequency cepstral coefficients[J]. variations, 2004, 1(4): 565–568
41. Manzo-Martinez A, Camarena-Ibarrola A (2011) A robust characterization of audio signals using the level of information content per Chroma[C]. 2011 IEEE International symposium on signal processing and information technology (ISSPIT). IEEE, 212–217
42. Harte C, Sandler M, Gasser M (2006) Detecting harmonic change in musical audio[C]. Proceedings of the 1st ACM workshop on audio and music computing multimedia. pp 21–26
43. Grosche P, Müller M, Kurth F (2010) Cyclic tempogram-A mid-level tempo representation for musicsignals[C]. 2010 IEEE International conference on acoustics, speech and signal processing. IEEE, pp 5522–5525
44. Müller M, Ewert S (2011) Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features[C]. Proceedings of the 12th International conference on music information retrieval (ISMIR), 2011. hal-00727791, version 2–22 Oct 2012
45. Grosche P, Müller M, Kurth F (2010) Cyclic tempogram-A mid-level tempo representation for musicsignals[C]. 2010 IEEE International conference on acoustics, speech and signal processing. IEEE, pp 5522–5525
46. Haitsma J, Kalker T (2003) A Highly Robust Audio Fingerprinting System With an Efficient Search Strategy. J New Music Res 32(2):211–221, 01 June 2003
47. Dapello J, Marques T, Schrimpf M et al (2020) Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations[J]. Adv Neural Inf Process Syst 33:13073–13087
48. Woo S, Park J, Lee JY et al (2018) Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). pp 3–19
49. La Grassa R, Gallo I, Landro N (2021) Learn class hierarchy using convolutional neural networks[J]. Appl Intell 51(10):6622–6632
50. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7263–7271
51. Sun M, Yuan Y, Zhou F et al (2018) Multi-attention multi-class constraint for fine-grained image recognition[C]. Proceedings of the European conference on computer vision (ECCV), pp 805–821
52. Srivastava DRK Pandey D (2022) Speech recognition using HMM and Soft Computing. Mater Today: Proc 51:1878–1883, 01 Jan 2022
53. Wijayasingha L Stankovic JA (2021) Robustness to noise for speech emotion classification using CNNs and attention mechanisms. Smart Health, 19:100165, 01 Mar 2021
54. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition, IEEE
55. Vaswani A, Shazeer N, Parmar N et al (2016) Attention is all you need[J]. Adv Neural Inform Process Syst 30
56. Available online: https://cis.whoi.edu/science/B/whalesounds/index.cfm
57. Cho K et al (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Comput Sci
58. Technicolor T, Related S, Technicolor T, Related S (2012) ImageNet Classification with Deep Convolutional Neural Networks [50]
59. Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. Comput Sci
60. Szegedy C, Liu W, Jia Y, Sermanet P, Rabinovich A (2014) Going Deeper with Convolutions. IEEE Computer Society
61. Available online: https://www.xeno-canto.org
62. Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research[C]. Proceedings of the 22nd ACM international conference on multimedia, pp 1041–1044

63. González-Hernández FR, Sánchez-Fernández LP, Suárez-Guerra S, Sánchez-Pérez (2017) Marine mammal sound classification based on a parallel recognition model and octave analysis. Appl Acoust 119:17–28, 01 April 2017
64. Trawicki MB (2021) Multispecies discrimination of whales (cetaceans) using Hidden Markov Models (HMMS). Ecol Inform 61:101223, 01 Mar 2021