引用格式：潘齐炜, 程吉祥, 田甜, 等. 基于特征融合与注意力机制的鸟类声纹识别方法[J]. 声学技术, 2024, **43**(5): 686-695. [PAN Qiwei, CHENG Jixiang, TIAN Tian, et al. Bird call recognition based on feature fusion and attention mechanism[J]. Technical Acoustics, 2024, **43**(5): 686-695.] **DOI**: 10.16300/j.cnki.1000-3630.2024.05.011

基于特征融合与注意力机制的鸟类声纹识别方法

潘齐炜，程吉祥，田 甜，吴 丹，曾 蕊
(西南石油大学电气信息学院，四川成都 610500)

摘要：鸟类声纹识别技术是一种将经过预处理的多种鸟类声音作为输入，通过网络模型识别出相应鸟类的技术。针对真实环境下鸟类声纹识别中单一音频特征局限和模型学习特征能力不佳问题，文章提出了一种基于特征融合和注意力机制的鸟类声纹识别方法。首先，在特征提取时分别获取梅尔频率倒谱系数和功率正则化倒谱系数，其次利用均值和方差归一化处理将两种特征融合得到新型融合特征参数MPFC；然后，以ResNet-50为主干网络在其残差模块中引入轻量化坐标注意力机制得到改进网络模型—坐标注意力残差网络；最后，将融合特征分别输入到坐标注意力残差网络(residual coordinate attention net, ResCA)，ResNet-50、ResNeSt-50、DenseNet-121和EfficientNet-B0并在两个数据集Birdsdata和BirdCLEF上进行对比实验。实验结果表明，融合特征比单一特征有更好的表征能力，能够提高一定识别率，改进网络也具有较好的识别效果。

关键词：鸟类声纹识别；特征融合；梅尔频率倒谱系数；功率正则化倒谱系

中图分类号：TN912.3 **文献标志码：**A **文章编号：**1000-3630(2024)-05-0686-10

Bird call recognition based on feature fusion and attention mechanism

PAN Qiwei, CHENG Jixiang, TIAN Tian, WU Dan, ZENG Rui
(*School of Engineering and Information*, *Southwest Petroleum University*, *Chengdu* 610500, *Sichuan*, *China*)

Abstract: Bird call recognition technology is a kind of technology that uses a variety of bird sounds as input after preprocessing, and identifies the corresponding bird species through the network model. In real natural environment, the single audio feature in bird call recognition has a limitation that the characteristics of bird calls cannot be fully described from preprocessing and the learning ability of the network model is poor. In this paper, a bird call recognition method based on feature fusion and attention mechanism is presented. First, Mel frequency cepstrum coefficients and power-normalized cepstral coefficients are obtained during feature extraction in the bird calls preprocessing stage. Secondly, the two features are fused by using the mean and variance normalization processing to obtain a new fusion feature called MPFC. Then, ResNet-50 is used as the backbone network, and by inserting coordinate attentionm mechanism into its residual module to improve the network model, an improved attention residual network model called ResCA can be obtained. Finally, the fusion features are respectively input to the ResCA, ResNet-50, ResNeSt-50, DenseNet-121 and EfficientNet-B0 for comparison in the two datasets Birdsdata and BirdCLEF. The results show that the fusion feature has better characterization ability than the single feature, and can improve the recognition rate. The improved network also has a better recognition effect.

Key words: bird call recognition; feature fusion; Mel frequency cepstrum coefficient; power-normalized cepstral coefficient

0 引 言

鸟类位于生态食物链的顶端，是反映环境是否健康的重要指标[1]。研究鸟类不是一项简单的任务，许多鸟类的栖息在与世隔绝难以到达的高海拔地区，研究者们只能通过长途跋涉了解当地鸟类受到环境变化和生态保护工作的影响。这给物理监测带来一定困难，科学家开始将监测重点转移至鸟类声音的监测上。目前在鸟类声音的监测上，鸟类生物学家的观察研究需要手动注释每个记录。这会耗费大量时间并需要专门的人员培训。机器学习的出现使人们可以通过大量的样本训练来自动识别不同鸟类的叫声。在机器学习的帮助下，研究人员能够

收稿日期: 2022-10-23; **修回日期:** 2023-02-24
基金项目: 国家自然科学基金(61603319,61601385)、西南石油大学智能控制与图像处理青年科技创新培育团队(2017CXTD010)。
作者简介: 潘齐炜(1998－), 男, 浙江永嘉人, 硕士生, 研究方向为声纹识别领域。
通信作者: 程吉祥, E-mail: chengjixiang0106@126.com

利用鸟类声音进行检测和分类，跟踪生态系统中的生物多样性情况以及变化趋势，更好地支持全球生态保护工作。

鸟类声纹识别技术将声纹识别应用到鸟类声音的识别中，首先将多种鸟类声音经过预处理后进行特征提取，然后与建立的声学模型进行匹配，将匹配值高的声音分类到相应的鸟类。随着人工智能的快速发展，学者们将深度学习技术引入声纹识别中，利用鸟类声纹进行自动识别分类的任务研究并取得了不错的进展[2]。本文主要提出了一种将(Mel frequency cepstrum coefficient, MFCC)特征与(power-normalized cepstral coefficient, PNCC)融合的方法和一种注意力机制改进ResNet-50的模型，在预处理和模型两个方面进行改进，进一步提高鸟类声纹的识别率。

1　声纹识别研究概况

1.1　声纹识别

声纹识别可以分为声纹确认和声纹辨认。声纹确认是判断输入声音是否为对比样本，类似于声音锁；声纹辨认是指判定测试声纹属于目标声纹集合中哪一种样本。鸟类声纹识别就属于声纹辨认的应用。

声纹识别技术大致可分为三个发展阶段[3]：在20世纪50年代，Kesta等首次提出声纹这一概念，但其仅利用人耳和肉眼进行识别操作，识别准确率较低。在20世纪末，Reynolds等[4]提出使用高斯混合模型(Gaussian mixture model, GMM)来拟合声纹模型，通过GMM能够尽可能模拟出多维矢量任意连续概率分布，成为了当时的主流技术，声纹识别技术进入了一个新阶段。Reynolds等[4]提出的高斯混合模型-通用背景模型(Gaussian mixture model-universal background model, GMM-UBM)以GMM方法为基础并进行了改进，通过构建大量非目标用户的声音背景数据库，将其混合起来充分训练一个GMM。进入21世纪后，为了克服GMM-UBM模型信道抗干扰性弱、高斯分量相互独立的局限性，文献[5-6]中先后提出了联合因子分析(joint factor analysis, JFA)和i-Vector模型。JFA的中心思想是将声音模型所在的空间进行划分，并将信道相关特征删去；i-Vector属于JFA的改进方法，将声音映射成一个固定低维向量。这也是声纹识别技术的第二个分水岭。之后深度学习在声纹识别、图像处理等领域里快速发展并得到成功应用，声纹识别技术也具有更广泛的应用前景。Li等[7]在2017年发布了Deep Speaker端到端神经说话人嵌入系统，建立了深度残差卷积神经网络模型(residual convolutional neural network, ResCNN)，使用深度神经网络提取语音信号中的帧级特征，再用池化层和长度归一化层产生特征信息通过损失函数进行训练。2021年Hourri等[8]通过实验证明了CNN可以直接应用于声纹识别，并且提出了ConvVector的新向量。相比较于传统方法，深度学习方法在处理任务上性能表现更加优秀。

1.2　鸟类声纹识别

GMM成为主流声纹识别技术之后，也被应用于鸟类声纹识别。Ptacek等[9]使用GMM-UBM模型用进行鸟类种类识别，虽然识别率不高，但也是对鸟类声纹识别的一次探究。在2008年，Lee等[10]通过提取鸟类声音信号通过提取MFCC，并利用GMM训练不同鸟类样本达到识别验证效果。在2014年，王恩泽等[11]提出一种基于MFCC特征参数合双重GMM模型的鸟类声纹识别方法，其识别精确度高于传统的单GMM模型。

在进入21世纪之后，机器学习的兴起将声纹识别技术引入第二个分水岭。Chakraborty等[12]在文献中提出使用了基于动态核的支持向量机SVM对喜马拉雅山下游地区26种鸟类完成分类识别，识别准确率不低于80%。基于深度学习方法的声纹识别研究从此开始。

相比于传统方法，基于深度学习的方法也受到了更加广泛的关注。2014年起，全球举办了多项围绕鸟类声纹识别技术展开的挑战赛，包括Bird-CLEF[13]、DCASE[14]等。近年的多项研究都使用了深度学习方法进行鸟类声音的分类与识别。Ming Zhong、Ruth Taylor的研究团队[15]为了研究尼泊尔地区2018—2019年生态系统的物种活动，现场记录黄喉莺和红喉鹪鹩叫声并使用卷积神经网络CNN实现有效分类。Efremava等[16]在2019年使用ResNet-50残差网络将有限数量的鸟类声音样本完成自动分类，将2 814个不同鸟类数据集样本归到了10个目标数据集种。2020年，杨春勇等[17]利用KNN算法和SVM结合，识别包含鸟类声纹信息的能量谱图，并使用生成对抗网络GAN进行图像增强来解决数据集样本不足的问题，其最高识别率可达到92%。Maegawa等[18]在2021年研究苍鹰生活环境以及繁殖情况，将记录的声纹数据转换成图片再通过卷积神经网络实现自动识别分类，总体准确率达到97%。这些实验数据表明，声纹识别技术在深度学习的加持下达到良好的效果，基于深度学习的

声纹识别技术成为发展趋势。

2 基于特征融合的声纹预处理

鸟类声纹的能量频谱图是研究鸟类声纹识别技术的一个重要途径，可以通过特征提取的方式生成包含相关信息的能量频谱图，显示有关鸟类声音能量分布的特性。能量频谱图具有频谱分布和时域波形的特点[19]。能量频谱中亮度越高代表该时段的鸟类鸣叫声能量越高，不同鸟类的能量频谱图存在差异，声纹识别可以利用这一特征识别鸟类。

声音是特殊的生物特征，具有一维性和易变性，识别时需要通过将声纹中可分性强且稳定性高的声学特征提取出来以降低识别的难度。因此声学特征的选择，对声纹识别率有很大影响。研究人员综合考虑人耳的听觉属性，通过加重低频段、降低高频段信号，均衡不同频段的能量差异以及除去系数相关性等方法，声纹中的原始信息得到有效保留，降低了识别难度，降低了训练数据的相对复杂性。特征提取是将原始音频信号以特征向量的形式表示并建模。从不同角度建模会生成不同的音频特征，从模拟人发声机理出发，得到线性预测分析特征(linear predictive coding, LPC)。将LPC换算到倒谱域便得到线性预测倒谱系数(linear prediction cepstral coefficients, LPCC)，LPCC对元音有较好的描述能力，但是抗噪性差。从模拟人耳听觉机理出发，得到感知线性预测系数(perceptual linear predictive, PLP)[20]，这样有利于抗噪语音特征信息的提取。最常使用的音频特征是梅尔频率倒谱系数(MFCC)，通过Mel尺度模拟人耳听觉的非线性，再利用三角滤波器进行特征提取即可得到MFCC特征[21]。MFCC的出现在一定程度上提高了识别率，但在有噪声的情况下，MFCC的识别效果也不理想。Kim等[22]提出的功率正则化倒谱系数PNCC，主要使用伽玛通(Gammatone)函数模拟人耳听觉响应，在噪声情况下具有一定鲁棒性。

2.1 MFCC特征提取

MFCC特征通过Mel频率尺度模拟人耳听觉频率的非线性。Mel频率与线性频率之间的关系能够近似表示为

$$f_{\text{Mel}}=2\,595\lg\left(1+\frac{f}{700}\right) \tag{1}$$

在MFCC特征提取过程中，对信号低频部分进行增强，减少扰动成分，以突出有效信息。MFCC特征提取流程如图1所示。

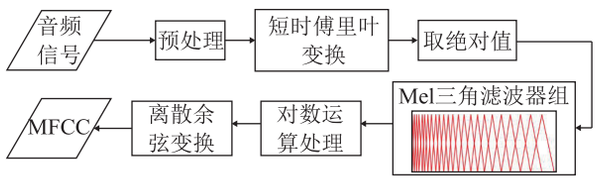


图1 MFCC特征提取流程图

Fig.1 Flowchart of MFCC feature extraction

输入的声音信号首先经过预加重、分帧、加窗以及短时傅里叶变换等预处理操作，将信号数据从时域转换到频域，然后取模平方得到谱线能量：

$$P_i(k)=\frac{1}{N}\left|S_i(k)\right|^2 \tag{2}$$

其中：N为分帧的大小，$S_i(k)$表示每帧音频数据，i为该帧音频的帧号。

将谱线能量使用M个Mel三角滤波器组进行滤波处理，将频谱平滑化，并得到系数$m_1,m_2,\cdots,m_M$。三角滤波器的频率响应定义为

$$H_{\text{m}}(k)=\begin{cases}0 & ,k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m)\leqslant k\leqslant f(m+1)\\ 0 & ,f(m+1)\leqslant k\end{cases} \tag{3}$$

其中：$\sum_{m=1}^{M}H_m(k)=1$,$f(m)$为中心频率，本文M为26。

谱线能量$P_i(k)$经过滤波器$H_m(k)$得到频谱能量$P_m(k)$：

$$P_m(k)=\sum_{m=f(m-1)}^{f(m+1)}H_m(k)P_i(k),m=0,1,2,\cdots,M-1 \tag{4}$$

对数运算操作能够很好地描述人耳听觉系统的非线性。将经过滤波的频谱能量进行对数运算，得到对应的对数能量：

$$S(m)=\ln\left(P_m(k)\right),0\leqslant m\leqslant M \tag{5}$$

将对数能量进行离散余弦变换，将对数能量变换到时域，得到了MFCC系数，表达式为

$$C(n)=\sum_{m=1}^{M}s(m)\cos\frac{\pi n(m-0.5)}{M} \tag{6}$$

其中：L表示MFCC系数的维度数，第1维表示能量参数，14~26维参数量过多且变化不大，通常取2~13维系数作为MFCC特征参数。

2.2 PNCC特征提取

功率正则化倒谱系数PNCC特征提取过程中增加了对音频短期功率和长时功率的处理，提高了噪声抗扰性，在混响环境下能够得到比MFCC更高的

识别率，如图2所示。

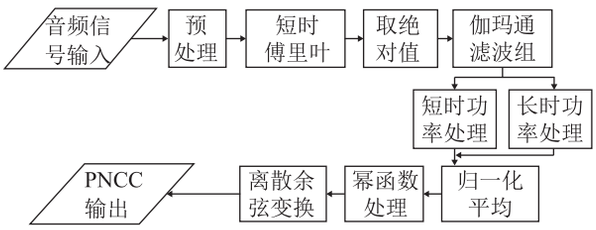


图2　PNCC特征提取流程图

Fig.2　Flowchart of PNCC feature extraction

PNCC中通常使用40维伽玛通(Gammatone)线性滤波器，用来模拟耳蜗频率分解特点，其幅频特性与人耳特性都是近似对称，其时域表达式为

$$h(t)=\begin{cases}0 & ,t>0\\ ct^{n-1}\exp(-2\pi bt)\cos(2\pi f_0 t+\phi)U(t), & t\leqslant 0\end{cases} \tag{7}$$

其中：c为滤波器增益系数，常见取值为0.01~10之间；n为滤波器阶数；f_0为滤波器中心频率；$U(t)$为阶跃信号；b为时间衰减因子，通常为0.9~1.0，b越大滤波时间越短；ϕ是滤波器的相位。

Gammatone滤波器组的响应在每个信道中的传递函数都满足：

$$\sum_{k=0}^{\left(\frac{K}{2}\right)-1}\left|H_l\left(\mathrm{e}^{\mathrm{j}\omega_k}\right)\right|^2=1 \tag{8}$$

其中：$H_l\left(\mathrm{e}^{\mathrm{j}\omega_k}\right)$是第$l$个信道的Gammatone滤波器在$\omega_k$处的响应，$K$表示短时傅里叶变换的长度。

将输入通过Gammatone滤波器组进行处理，然后进行短时功率处理与长时功率处理。短时频功率定义为

$$P(m,l)=\sum_{k=0}^{\frac{k}{2}-1}\left|X\left(m,\mathrm{e}^{\mathrm{j}\omega_k}\right)H_l\left(\mathrm{e}^{\mathrm{j}\omega_k}\right)\right|^2 \tag{9}$$

其中：m表示输入音频帧数，l表示滤波器通道数。

由于噪声能量变化没有音频变化快。通过长时功率处理对每一帧噪声进行平滑化能够减少噪声干扰，根据式(10)得到频带m处的功率平滑值$\bar{Q}(m,l)$用于之后的噪声估计和补偿，具体表达式为

$$\bar{Q}(m,l)=\frac{1}{2M+1}\sum_{m'=m-M}^{m+M}P(m,l) \tag{10}$$

式中：选择$M=2$时对性能的影响最小。

接下来利用非对称噪声抑制和临时掩蔽滤波处理，过滤掉信号低频部分减弱混响，获得$\bar{R}[m,l']$。在不同通道之间做平滑处理，进行光谱权重平滑化，得到归一化的频带能量比值的平滑结果$\bar{S}(m,l)$进一步降低时间上的噪声影响，使得特征提取具有更强的鲁棒性：

$$\bar{S}(m,l)=\frac{1}{l_2-L_1+1}\sum_{l'=l_1}^{l_2}\frac{\bar{R}(m,l')}{\bar{Q}(m,l)} \tag{11}$$

对$P(m,l)$与$\bar{S}(m,l)$进行能量平滑处理得到经过归一化和平滑调整后的能量特征$T(m,l)$：

$$T(m,l)=P(m,l)\bar{S}(m,l) \tag{12}$$

使用能量归一化可减少在PNCC特征参数提取过程中振幅缩放的潜在影响，利用平均功率估计差分方程形式进行归一化：

$$\mu(m)=\lambda_\mu\mu(m-1)+\frac{\left(1-\lambda_\mu\right)}{L}\sum_{l=0}^{L-1}T(m,l) \tag{13}$$

其中：平滑因子λ_μ取值为0.999，L表示通道总数。

接着进行功率归一化调整，消除直流功率的干扰，得到归一化后的能量特征$U(m,l)$：

$$U(m,l)=k\frac{T(m,l)}{\mu(m)} \tag{14}$$

其中：k为任意实数。

与MFCC特征参数不同，PNCC特征参数提取时在非线性处理部分采用幂函数进行归一化。幂函数存在阈值效应，当输入很小时，幂函数对应的输出约为零，这也很符合人耳的听觉特性，其表达式为

$$V(m,l)=U(m,l)^{\frac{1}{15}} \tag{15}$$

其中：声压指数为$\frac{1}{15}$的幂函数曲线与人的生理数据最为接近，$V(m,l)$表示对$U(m,l)$进行非线性处理后的特征参数。

最后经过离散余弦变换，可以得到PNCC特征参数。

2.3　特征融合

单一特征参数不能完全表征鸟类声音的所有特点，存在一定的局限。2018年仲伟锋等[23]为了进一步提高说话人识别准确性能融合浅度和深度特征进一步全面说话人的声纹信息。Jiang的研究团队[24]在2019年提出将MFCC特征和Mel频谱融合用于语音情感识别，能够比单特征识别具有更高的分类精度。虽然已有的一些研究使用多种特征融合方法提高了语音识别的准确性，但并不适用于鸟类声纹识别领域。因此针对鸟类声纹识别单一音频特征对声音表征的局限性，本文提出了将MFCC特征和PNCC特征进行拼接融合的方法，得到一种新的特征——MPFC特征。增强了对音频特征的表征能力，从而提高识别准确率，并为鸟类声纹识别研究提供一种新的思路。

本文提出的特征融合方法主要是将MFCC特征信息和PNCC特征信息进行拼接融合。特征融合流程图如图3所示。利用PNCC良好的抗噪声鲁棒性来填补MFCC抗噪性能的不足，MFCC能很好描

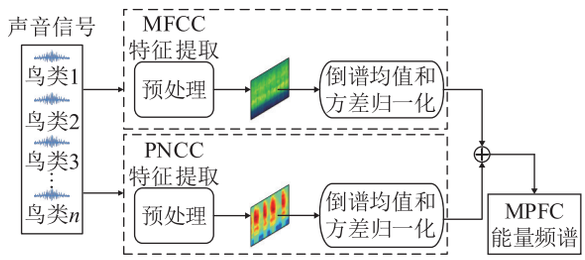
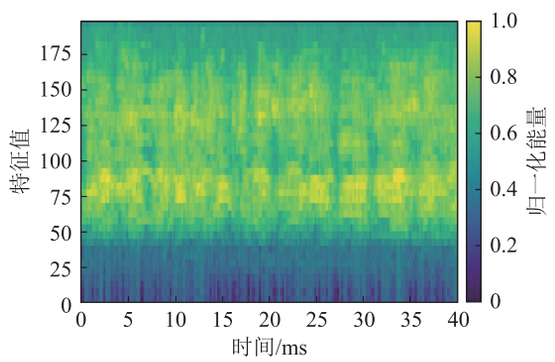


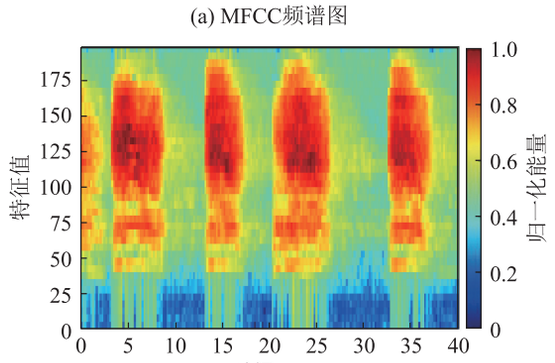
图3 特征融合流程图

Fig.3 Flow chart of feature fusion

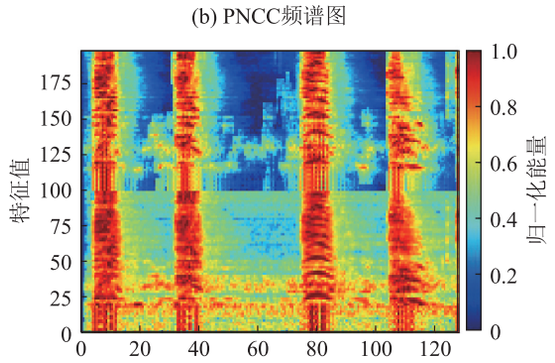
述一定范围内频谱的高频结构并忽略低频段，与PNCC形成一定的互补。为了减少信号失配问题将MFCC特征与PNCC特征进行倒谱均值和方差归一化处理，再将MFCC和PNCC进行拼接融合得到新型特征MPFC。某段鸟类声纹能量频谱图示例如图4所示。



(a) MFCC频谱图



(b) PNCC频谱图



(c) MPFC频谱图

图4 某段音频的不同特征能量的声谱图

Fig.4 Spectrograms of different characteristic energies of a certain audio segment

3 注意力残差网络ResCA

3.1 模型总体结构

本文提出一种基于注意力机制改进的注意力残差网络(ResCA)结构。该网络由16个残差注意力单元、全局平均池化层和Softmax分类层构成，网络结构如图5所示。由于输入的频谱图分辨率相对较大，在该网络中将使用7×7卷积层和最大池化层对输入频谱图进行下采样，224×224的分辨率降至112×112，尽量保留原始图像细节再输入至注意力残差单元中。

原残差网络(ResNet-50)结构一共有3+6+4+3=16个瓶颈层(bottleneck)，每个瓶颈层依次由1×1卷积层、3×3卷积层、1×1卷积层与短路连接构成，将瓶颈层的残差模块更换为注意力残差单元，再经过全局平均池化层、全连接层、Softmax层输出分类结果。

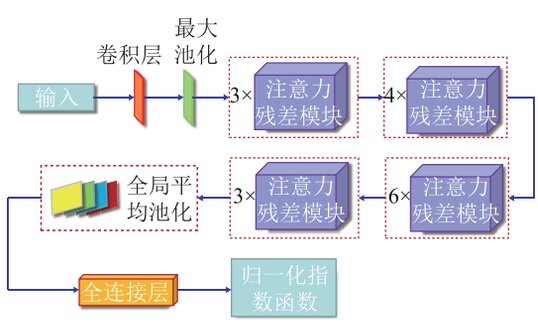


图5 ResCA网络结构图

Fig.5 Structure diagram of ResCA network

3.2 注意力残差模块

ResCA网络包含16个由残差模块改进的注意力残差模块。注意力残差模块由一条残差网络模块支路、坐标注意力以及短路连接组成。输入特征依次经过1×1卷积层、3×3卷积层、1×1卷积层去除冗余信息以降低特征维度，然后通过注意力机制突出局部区域与直接经过卷积的部分叠加，能够捕捉有用的特征信息以提升识别率。由于外加短路连接，也不会因为没获取到重要信息而降低网络学习能力，注意力残差模块的结构图如图6所示。

注意力机制是深度学习中的一项重要技术[25]，近年来在目标监测、图像分割以及自然语言处理等领域取得了突破，有效提高了模型的性能。其主要思想是基于人类视觉特有的大脑信号处理机制。人类快速浏览整体图像时会获取自身需要的重点局部信息，而自动忽略相对低价值、无用的信息。利用这样的机制，能够快速提取对自己有利的信息，提

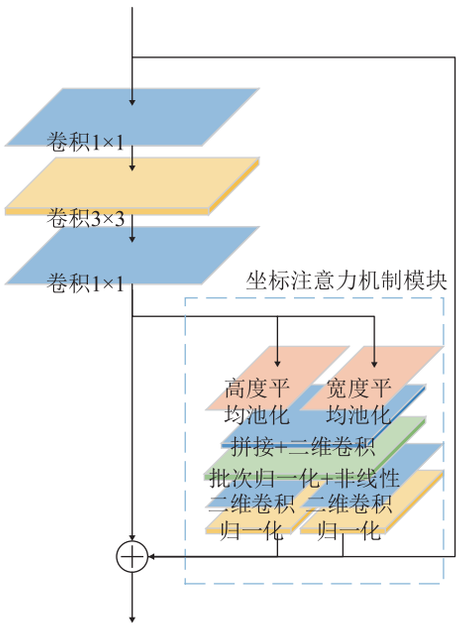


图6　注意力残差单元结构图

Fig.6　Structure diagram of attention residual unit

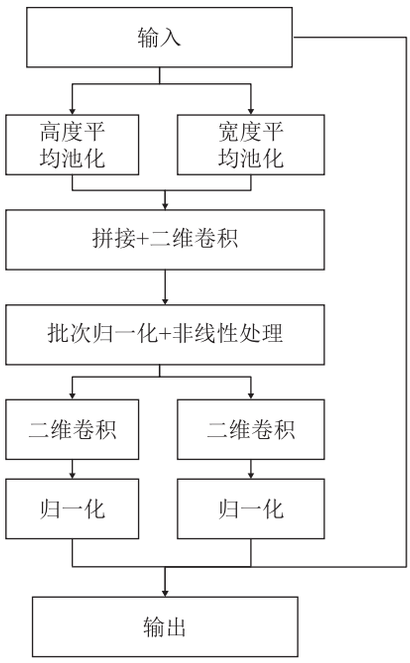


图7　坐标注意力机制结构图

Fig.7　Structure diagram of the coordinated attention mechanism

高信息处理的有效性。最早的Encoder-Decoder框架[26]，在计算能力有限的情况下，优先从模型大量参数信息中筛选出对当前任务更为关键有用的信息。最经典的是SENet模型。由Momenta公司[27]在2017年发布的SENet模型是将每个二维特征进行压缩，通过二维全局池化将特征转换为单个特征向量，建立通道之间的相互依赖关系。

本文使用坐标注意力机制改进(coordinate attention) ResNet-50。相比SENet该机制增加了特征空间位置的信息，提高了特征中重要信息的利用率。2021年Hou等[28]提出一种轻量级网络设计的注意力机制coordinate attention，与通道注意力不同的是为了减少二维全局池化造成的位置信息丢失，coordinate attention将通道注意力分解为两个并行的一维特征解码过程，生成的特征信息能够互补重点信息并将其应用到特征图中。该结构能够灵活插入网络中，如图7所示。

给定输入向量$\boldsymbol{x}$，分别使用$(H,1)$和$(1,W)$的池化核沿着水平和垂直方向对第c个通道进行编码，则h高度第c个通道输出z_c^h。同理宽度为w、第c个通道的输出z_c^w，表示为

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,i) \tag{16}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,w) \tag{17}$$

根据式(16)、(17)得到$z_c^h(h)$和$z_c^w(w)$并生成一对方向感知特征图，从特征图中能够获得全局感受野且编码精确的位置信息。再使用一个1×1的卷积操作F_1进行变换，得到包含位置信息的中间特征f：

$$f=\delta\left\{F_1\left(\left[z^h\ z^w\right]\right)\right\} \tag{18}$$

将中间特征图f进行拆分，经过1×1的卷积操作F_h、F_w，得到g^h和g^w，作为注意力权重：

$$g^h=\sigma\left[F_h\left(f^h\right)\right] \tag{19}$$

$$g^w=\sigma\left[F_w\left(f^w\right)\right] \tag{20}$$

最终得到坐标注意力模块的输出y，可以表示为

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{21}$$

本文提出的坐标注意力残差网络ResCA网络配置如表1所示。

表1　ResCA网络配置参数表

Table 1　ResCA network configuration parameters

网络层名称	输出尺寸	50层网络
Conv1	112×112	7×7, stride=2
Conv2_x	56×56	3×3 MaxPooling $\begin{bmatrix}1\times1 & 64\\ 3\times3 & 64\\ 1\times1 & 256\end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix}1\times1 & 128\\ 3\times3 & 128\\ 1\times1 & 256\end{bmatrix}\times4$
Conv4_x	14×14	$\begin{bmatrix}1\times1 & 128\\ 3\times3 & 256\\ 1\times1 & 1024\end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix}1\times1 & 512\\ 3\times3 & 512\\ 1\times1 & 2048\end{bmatrix}\times3$
	1×1	AvePooling, Fc, Softmax
计算量		3.6×10^9

4 实验与分析

4.1 鸟类声纹数据集

实验使用Birdsdata和BirdCLEF挑战赛[13]的两个数据集。Birdsdata是国内开源数据集(http://www.birdsdata.com/english)，是由北京智源人工智能研究院和百鸟数据联合发布的自然声音检测数据集，包含了灰雁、大天鹅、绿头鸭、绿翅鸭、灰山鹑、西鹌鹑、雉鸡、红喉潜鸟、苍鹭、普通鸬鹚、苍鹰、欧亚鸲、西方秧鸡、骨顶鸡、黑翅长脚鹬、凤头麦鸡、白腰草鹬、红脚鹬、林鹬、麻雀共20种鸟类、14 311个鸟类录音样本。该数据集的声音数据都是在自然场景中收集，主要用于自然界鸟类采集声音类别的检测。BirdCLEF挑战赛数据集来自xenocanto.org网站上的397种鸟类的简短录音，本文中再从中随机选取50种鸟类简短录音作为自制数据集，一共5 955个鸟类录音样本。

4.2 实验设计

本文所使用数据集声音均保持统一参数设置：采样率为16 kHz，帧长为25 ms，帧移为2 ms。实验过程中将音频数据转换得到的能量频谱图裁剪为224×224个像素，并随机分为训练集、测试集和验证集，占比分别为0.6、0.2与0.2。

首先对声音信号进行预处理，$H(z)=1-\lambda z^{-1}$，取$\lambda=0.9$，然后进行分帧并加汉明窗，最后分别提取MFCC、PNCC以及MPFC三种特征，并利用ResNet-50网络进行实验效果对比。为验证本文的注意力残差网络ResCA的有效性，将融合特征MPFC能量频谱图作为输入，在数据样本一致的情况下将ResCA网络与四种网络模型ResNet-50、ResNeSt-50[29]、DenseNet-121[30]、EfficientNet-B0[31]进行对比实验。

本文所提方法在Visual Studio Code软件上实现，使用Pytorch作为深度学习框架。计算机操作系统为Windows10 64bit，硬件配置为8G显存NVIDIA GeForce RTX2080ti GPU，3.6GHz Intel Core CPU，i5-10600处理器。

实验使用准确率、错误率、混淆矩阵和F1score作为评价指标。准确率(R_{Acc})是指真实值与预测值相同的样本占总样本数的比例，错误率(R_{Err})为不相同的样本数占总数比例。F1Score为精确度(P)和召回率(R_{REC})的平均值，其公式为

$$F_{1\mathrm{Score}}=2\times\frac{P\times R_{\mathrm{Rec}}}{P+R_{\mathrm{Rec}}} \tag{22}$$

4.3 实验结果

4.3.1 总体实验结果比较

在网络模型对比实验中，将每个网络模型训练迭代次数设为300，尽可能使每个网络模型能够学习到足够特征信息，在两个数据集中得到最好的结果。表2显示了五种网络模型在两个数据集下的分类准确率。从表2数据中看出，Birdsdata数据集中得到指标均比自制BirdCLEF数据集中的表现要好。自制BirdCLEF数据集的音频是由xenocanto.org网站从全世界鸟类研究者或鸟类爱好者处收集的，音频中存在的噪声混响等干扰可能比Birdsdata数据集多，以及选取鸟的种类比Birdsdata数据集更多，也是前者指标比后者低的原因。坐标注意力残差网络ResCA在Birdsdata上的准确率均值为93.60%，F1score为90.17%。在自制BirdCLEF数据集中的准确率均值为72.12%，F1score为77.33%。DenseNet-121与EfficientNet-B0在两个数据集中的准确率都低于ResCA。对于模型ResNet-50在两个数据集上的性能表现，注意力残差网络ResCA准确率得到一定提高，即使ResCA的参数量稍有增加，但低于ResNeSt-50网络模型。相比另外四种网络模型，本文网络模型ResCA在数据集中的识别准确率较好。

图8显示了5个网络模型在自制BirdCLEF数据集下(随机选取数据集中10种不同鸟类)鸟类声音识别准确率。不难看出，坐标注意力残差网络ResCA模型的分类识别准确率大部分在50%以上，整体准

表2 在2个数据集中5种网络模型的识别结果

Table 2 Recognition results of the five network models in the two datasets

网络结构	Birdsdata数据集		BirdCLEF		参数量/10^6	计算量/10^9
	准确率/%	F1score/%	准确率/%	F1score/%		
ResNet-50	93.43	89.44	71.81	76.02	25.5	3.53
ResNeSt-50	93.66	90.34	72.56	73.88	30.5	3.8
DenseNet-121	90.62	90.59	70.17	72.53	7.3	2.79
Efficient-B0	89.10	89.15	67.77	65.94	5.3	3.9
ResCA	93.60	90.17	72.12	77.33	26.1	3.6

确率高于ResNet-50、DenseNet-121和EfficientNet-B0，仅低于ResNeSt-50模型。图9为在Birdsdata数据集5个网络模型的损失函数曲线，图中损失函数随着迭代次数的增加逐渐降低并趋于平稳收敛。坐标注意力残差网络ResCA模型的曲线斜率小于ResNeSt-50和DenseNet-121的loss曲线数下降表现，但是大于ResNet-50和EfficientNet-B0，收敛速度快于ResNet-50和EfficientNet-B0。ResCA对数据集分类的准确率有一定提升，加入坐标注意力机制后模型参数量没有大幅增加。

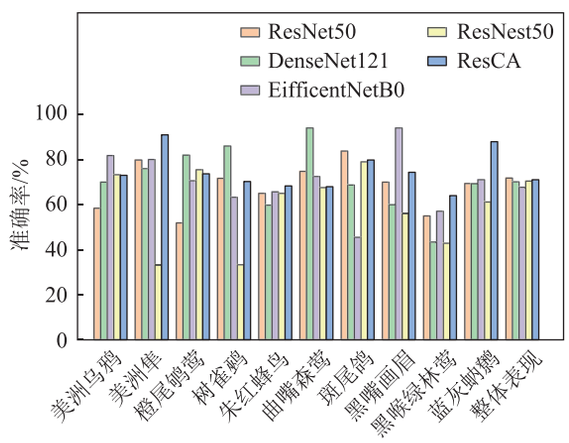


图8　在BirdCLEF数据集中5种网络模型识别10种鸟类声音的准确率

Fig.8　Recognition accuracy rates of ten birds voices for the five network models in the BirdCLEF dataset

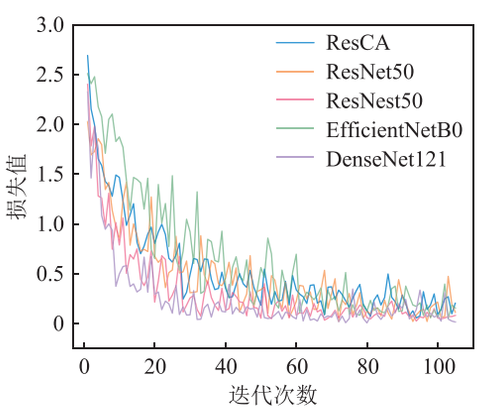
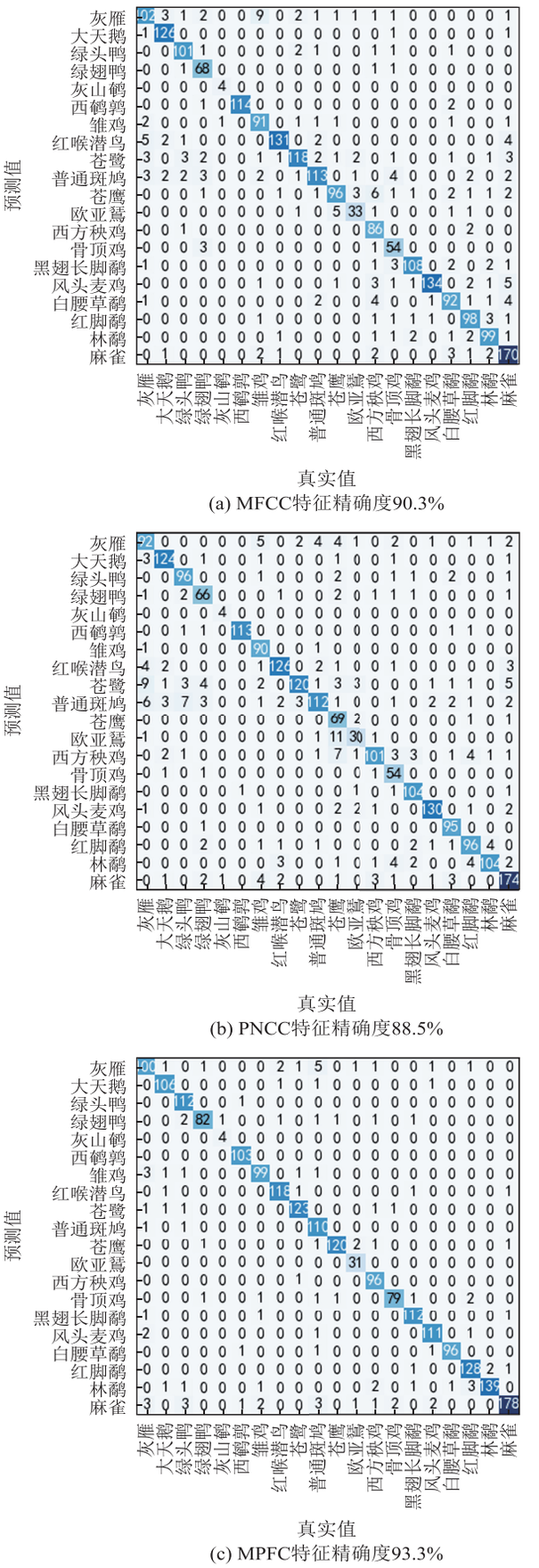


图9　在Birdsdata数据集中5种网络模型的损失曲线

Fig.9　Loss curves of the five network models in the Birdsdata dataset

4.3.2　特征融合有效性分析

MFCC、PNCC和MPFC三种特征在Birddata数据集中识混淆矩阵如图10所示，在ResNet-50网络模型中的精确度如图11所示。表3所示为三种音频特征的准确率和F1Score。由实验结果可得出，三种音频特征参数在两个不同的数据集中的表现，MPFC融合特征识别的精确度和准确率相比MFCC特征和PNCC特征识别的精确度、准确率都有较大



(a) MFCC特征精确度90.3%

(b) PNCC特征精确度88.5%

(c) MPFC特征精确度93.3%

图10　在Birdsdata数据集中3种特征的混淆矩阵

Fig.10　The confusion matrices of the three features in the Birdsdata dataset

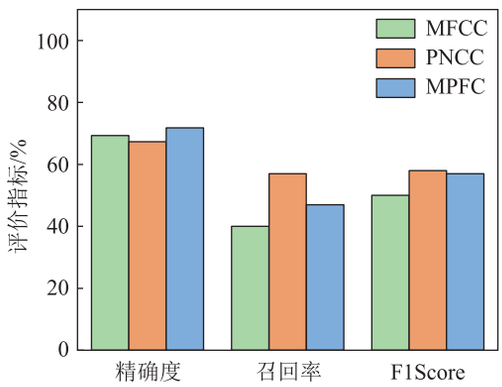


图 11 3 种特征在 BiredCLEF 数据集的评价指标

Fig.11 Evaluation indicators of the three features in the Bired-CLEF dataset

表 3 在 2 个数据集中 3 种特征的准确率和 F1Score 值

Table 3 Accuracy rates and F1Score values of the three features in the two datasets

特征类型	Birdsdata		BirdCLEF	
	准确率/%	F1Score/%	准确率/%	F1Score/%
MFCC	90.26	44.71	70.06	50.06
PNCC	88.49	43.31	69.14	60.14
MPFC	93.32	45.88	71.81	56.12

的提高，说明了 MPFC 融合特征参数在自然环境下的鸟类声纹识别中，精确度和识别率比单一特征参数的更好。

表 4 内容中展示了在 BirdCLEF 数据集种本文与文献[32]的 F1Score 指标。实验结果显示，在使用相同网络模型训练的情况下，MPFC 融合特征参数的 F1Score 高于 MFCC 特征参数识别 的 F1Score。

综合 BirdCLEF 数据集和 Birdsdata 数据集上的表现来看，本文提出的特征融合 MPFC 与坐标注意力残差网络 ResCA 表现相比较于 MFCC 特征、PNCC 特征以及四种网络模型具有更好的识别准确率。

表 4 本文方法与文献[32]中 3 种网络模型的 F1Score 值

Table 4 F1Score values of the three network models obtained in this paper and the literature [32]

网络模型	F1Score/%	
	文献[32]方法	本文方法
ResNeSt-50	70.60	73.88
DenseNet-121	66.00	72.53
Efficient-B0	69.10	65.94

5 结 论

本文提出了一种特征融合和坐标注意力残差网络结构的鸟类声纹识别方法。该方法通过将 MFCC 特征与 PNCC 特征进行融合拼接得到 MPFC 特征，增强了音频特征的表征能力，增大不同鸟类声纹之间的差异，提高了识别率。该方法同时利用坐标注意力机制改进 ResNet-50 网络结构，提高了对图像特征局部重点信息的提取能力，加快网络的收敛。在 Birdsdata 和 BirdCLEF 数据集中的实验结果证明本文的方法优于其他几种方法。本文方法还可进一步改进，例如将两种以上的音频特征组合、更好的网络结构改进等。在不同环境中的鸟类识别率也是今后的研究发展方向。

参 考 文 献

[1] XIE J, HU K, ZHU M Y, et al. Data-driven analysis of global research trends in bioacoustics and ecoacoustics from 1991 to 2018[J]. Ecological Informatics, 2020, **57**: 101068.

[2] XIE J, ZHONG Y, ZHANG J, et al. A review of automatic recognition technology for bird vocalizations in the deep learning era[J]. Ecological Informatics, 2022: 101927.

[3] BAI Z X, ZHANG X L. Speaker recognition based on deep learning: an overview[J]. Neural Networks, 2021, **140**: 65-99.

[4] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, **10**(1-3): 19-41.

[5] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, **19**(4): 788-798.

[6] DEHAK N, DUMOUCHEL P, KENNY P. Modeling prosodic features with joint factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, **15**(7): 2095-2103.

[7] LI C, MA X K, JIANG B, et al. Deep speaker: an end-to-end neural speaker embedding system[EB/OL]. 2017: arXiv: 1705.02304. https://arxiv.org/abs/1705.02304.

[8] HOURRI S, NIKOLOV N S, KHARROUBI J. Convolutional neural network vectors for speaker recognition[J]. International Journal of Speech Technology, 2021, **24**(2): 389-400.

[9] PTACEK L, MACHLICA L, LINHART P, et al. Automatic recognition of bird individuals on an open set using as-is recordings[J]. Bioacoustics, 2016, **25**(1): 55-73.

[10] LEE C H, HAN C C, CHUANG C C. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, **16**(8): 1541-1550.

[11] 王恩泽, 何东健. 含基于 MFCC 和双重 GMM 的鸟类识别方法[J]. 计算机工程与设计, 2014, **35**(5): 1868-1871.
WANG Enze, HE Dongjian. Bird recognition based on MFCC and dual-GMM[J]. Computer Engineering and Design, 2014, **35**(5): 1868-1871.

[12] CHAKRABORTY D, MUKKER P, RAJAN P, et al. Bird call identification using dynamic kernel based support vector machines and deep neural networks[C]//2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). Anaheim, CA, USA. IEEE, 2017: 280-285.

[13] JOLY A, GOËAU H, KAHL S, et al. Overview of LifeCLEF 2020: A System-Oriented Evaluation of Automated Species Identification and Species Distribution Prediction[C]//International Conference of the Cross-Language Evaluation Forum for European Languages. Cham: Springer, 2020: 342-363.

[14] de BENITO-GORRON D, RAMOS D, TOLEDANO D T. A

multi-resolution CRNN-based approach for semi-supervised sound event detection in DCASE 2020 challenge[J]. IEEE Access, 2021, **9**: 89029-89042.

[15] ZHONG M, TAYLOR R, BATES N, et al. Acoustic detection of regionally rare bird species through deep convolutional neural networks[J]. Ecological Informatics, 2021, **64**: 101333.

[16] EFREMOVA D B, SANKUPELLAY M, KONOVALOV D A. Data-efficient classification of birdcall through convolutional neural networks transfer learning[C]//2019 Digital Image Computing: Techniques and Applications (DICTA). Perth, WA, Australia. IEEE, 2020: 1-8.

[17] 杨春勇, 祁宏达, 彭焱秋, 等. 融合声纹信息的能量谱图在鸟类识别中的研究[J]. 应用声学, 2020, **39**(3): 453-463.
YANG Chunyong, QI Hongda, PENG Yanqiu, et al. Research on the application of energy spectrum with voiceprint information in bird recognition[J]. Journal of Applied Acoustics, 2020, **39**(3): 453-463.

[18] MAEGAWA Y, USHIGOME Y, SUZUKI M, et al. A new survey method using convolutional neural networks for automatic classification of bird calls[J]. Ecological Informatics, 2021, **61**: 101164.

[19] ZHANG X, CHEN A B, ZHOU G X, et al. Spectrogram-frame linear network and continuous frame sequence for bird sound classification[J]. Ecological Informatics, 2019, **54**: 101009.

[20] CHELALI F, DJERADI A, DJERADI R. Speaker identification system based on PLP coefficients and artificial neural network[C]//Proceedings of the World Congress on Engineering, 2011, 1: 1641-1646.

[21] DUFOUR O, ARTIERES T, GLOTIN H, et al. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification[J]. Soundscape Semiotics—Localization and Categorization, 2013 (2013): 89-93.

[22] KIM C, STERN R M. Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, Japan. IEEE, 2012: 4101-4104.

[23] 仲伟峰, 方祥, 范存航, 等. 深浅层特征及模型融合的说话人识别[J]. 声学学报, 2018, **43**(2): 263-272.
ZHONG Weifeng, FANG Xiang, FAN Cunhang, et al. Fusion of deep shallow features and models for speaker recognition[J]. Acta Acustica, 2018, **43**(2): 263-272.

[24] JIANG C, MAO R, LIU G, et al. Speech Emotion Recognition based on Multiple Feature Fusion[C]//2019 Chinese Automation Congress (CAC). IEEE, 2019: 907-912.

[25] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: a survey[J]. Computational Visual Media, 2022, **8**(3): 331-368.

[26] JI Y, ZHANG H, ZHANG Z, et al. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances[J]. Information Sciences, 2021, **546**: 835-857

[27] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 7132-7141.

[28] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[29] ZHANG H, WU C R, ZHANG Z Y, et al. ResNeSt: split-attention networks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, LA, USA. IEEE, 2022: 2735-2745.

[30] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 2261-2269.

[31] TAN M X, LE Q V. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.

[32] CONDE M V, SHUBHAM K, AGNIHOTRI P, et al. Weakly-supervised classification and detection of bird sounds in the wild. A BirdCLEF 2021 solution[EB/OL]. 2021: arXiv: 2107.04878. https://arxiv.org/abs/2107.04878.