PANNs：大规模预训练音频神经网络

音频模式识别

孔秋强，IEEE学生会员，曹寅，IEEE会员，Turab Iqbal，王宇轩，王文武，IEEE高级会员，Mark D.Plumbley，IEEE研究员

摘要音频模式识别是机器学习领域的一个重要研究课题，包括音频标记、声学场景分类、音乐分类、语音情感分类和声音事件检测等多项任务。最近，神经网络已被应用于解决音频模式识别问题。然而，以前的系统是建立在持续时间有限的特定数据集上的。最近，在计算机视觉和自然语言处理中，对大规模数据集进行预训练的系统已经很好地推广到了几个任务中。然而，关于大规模数据集上用于音频模式识别的预训练系统的研究有限。本文提出了在大规模AudioSet数据集上训练的预训练音频神经网络（PANNs）。这些PANN被转移到其他与音频相关的任务中。我们研究了由各种卷积神经网络建模的PANN的性能和计算复杂度。我们提出了一种称为Wavegram-Logmel-CNN的架构，该架构使用log-mel频谱图和波形作为输入特征。我们最好的PANN系统在AudioSet标记上实现了最先进的平均精度（mAP）0.439，优于之前最好的0.392。我们将PANN转移到六个音频模式识别任务中，并在其中几个任务中展示了最先进的性能。我们已经发布了PANN的源代码和预训练模型：https://github.com/qiuqiangkong/audioset_tagging_cnn.

索引术语音频标记、预训练音频神经网络、迁移学习。

一.导言

音频模式识别是机器学习领域的一个重要研究课题，在我们的生活中起着重要作用。我们被声音包围着，这些声音包含了我们所处位置的丰富信息，以及我们周围发生的事件。音频模式识别包含几个任务，如音频标记[1]、声学场景分类[2]、音乐分类[3]、语音情感分类和声音事件检测[4]。

近年来，音频模式识别引起了越来越多的研究兴趣。早期音频模式识别工作

Q.Kong、Y.Cao、T.Iqbal和M.D.Plumbley在英国Guild-ford GU2 7XH萨里大学视觉、语音和信号处理中心工作（电子邮件：q.kong@surrey.ac.uk；yin.cao@surrey.ac.uk；t.iqbal@surrey.ac.uk；m.plumbley@surrey.ac.uk）。

这项工作部分得到了EPSRC赠款EP/N014111/1《理解声音》的支持，部分得到了中国国家留学基金管理委员会研究奖学金201406150082的支持，另一部分得到了EP/N509772/1赠款下EPSRC博士培训伙伴关系的助学金（参考号：1976218）的支持。本研究得到了国家自然科学基金（11804365）的资助。（孔秋强为第一作者。）（尹曹为通讯

Y.Wang就职于美国加利福尼亚州山景城字节跳动人工智能实验室（电子邮件：wangyuxuan.11@bytedance.com）。

W.Wang来自英国吉尔福德GU2 7XH萨里大学视觉、语音和信号处理中心，以及中国青岛科技大学，邮编266071（电子邮件：w.wang@surrey.ac.uk）.

专注于个人研究人员收集的私人数据集[5][6]。例如，Woodard[5]应用隐马尔可夫模型（HMM）对三种类型的声音进行分类：木门打开和关闭、金属掉落和倒水。最近，声学场景和事件的检测和分类（DCASE）挑战系列[7][8][9][2]提供了公开可用的数据集，如声学场景分类和声音事件检测数据集。DCASE挑战引起了人们对音频模式识别越来越多的研究兴趣。例如，最近的DCASE 2019挑战赛在五个子任务中收到了311个参赛作品[10]。

然而，当在大规模数据集上训练时，音频模式识别系统的性能如何仍然是一个悬而未决的问题。在计算机视觉中，已经使用大规模ImageNet数据集构建了几个图像分类系统[11]。在自然语言处理中，已经使用维基百科等大规模文本数据集构建了几种语言模型[12]。然而，在大规模音频数据集上训练的系统更为有限[1][13][14][15]。

音频模式识别的一个里程碑是AudioSet的发布[1]，这是一个包含527个声音类别的5000多小时录音的数据集。AudioSet没有发布原始录音，而是发布了从预训练卷积神经网络中提取的音频片段的嵌入特征[13]。一些研究人员研究了具有这些嵌入特征的建筑系统[13][16][17][18][19][20]。然而，嵌入特征可能不是音频记录的最佳表示，这可能会限制这些系统的性能。在这篇文章中，我们提出了使用各种神经网络在原始AudioSet录音上训练的预训练音频神经网络（PANN）。我们发现，几个PANN系统的性能优于以前最先进的音频标记系统。我们还研究了PANN的音频标记性能和计算复杂性。

我们建议将PANN转移到其他音频模式识别任务中。之前的研究人员已经研究了音频标记的迁移学习。例如，在[21]中提出的百万首歌曲数据集上对音频标记系统进行了预训练，从预训练的卷积神经网络（CNN）中提取的嵌入特征被用作第二阶段分类器的输入，如在MagnaTagATune上预训练的神经项[23]，声学场景[24]数据集在其他音频标记任务上进行了微调[25][26]。这些迁移学习系统主要使用音乐数据集进行训练，并且仅限于比AudioSet更小的数据集。

networks or support vector machines (SVMs) [14][22]. Sys-

这项工作的贡献包括：（1）我们介绍

在AudioSet上训练的PANN有190万个音频片段，包含527个声音类别的本体；（2） 我们研究了各种PANN的音频标记性能和计算复杂度之间的权衡；（3） 我们提出了一个我们称之为Wavegram-Logmel CNN的系统，该系统在AudioSet标记上的平均精度（mAP）为0.439，优于之前最先进的mAP 0.392系统和谷歌的mAP 0.314系统；（4） 我们证明，PANN可以转移到其他音频模式识别任务中，优于几种最先进的系统；（5） 我们已经发布了源代码和预训练的PANN模型。

本文的结构如下：第二节介绍了使用各种卷积神经网络的音频标记；第三节介绍了我们提出的Wavegram CNN系统；第四节介绍了我们的数据处理技术，包括AudioSet标签的数据平衡和数据增强；第六节显示了实验结果，第七节总结了这项工作。。

二、音频标签系统

音频标记是音频模式识别的一项重要任务，其目标是预测是否存在。音频剪辑中的音频标签。音频标记的早期工作包括使用手动设计的特征作为输入，如音频能量、过零率和梅尔频率倒谱系数（MFCC）[27]，生成模型，包括高斯混合模型（GMMs）[28][29]、隐马尔可夫模型（HMM）和判别支持向量机（SVM）[30]已被用作分类器。最近，基于神经网络的方法，如卷积神经网络（CNN），已被用于预测录音的标签[3]。基于CNN的系统在几个DCASE挑战任务中取得了最先进的性能，包括声学场景分类[2]和声音事件检测[4]。然而，这些作品中的许多都集中在声音类别数量有限的特定任务上，并且不是为了识别各种各样的声音类别而设计的。在这篇文章中，我们专注于在AudioSet[1]上训练大规模PANNs，以解决一般的音频标记问题。

A. CNN

1） 传统的卷积神经网络：卷积神经网络已成功应用于计算机视觉任务，如图像分类[31][32]。CNN由几个卷积层组成，每个卷积层包含几个与输入特征图卷积的核，以捕获它们的局部模式。用于音频标记的CNN[3][20]通常使用log-mel频谱图作为输入[3][20]。短时傅里叶变换（STFT）应用于时域波形以计算频谱图。然后，将梅尔滤波器组应用于频谱图，然后进行对数运算以提取对数梅尔频谱图[3][20]。

2） 使CNN适应AudioSet标签：我们使用的PANN基于我们之前为DCASE 2019挑战提出的时间任务CNN系统[33]，并在CNN的倒数第二层添加了一个额外的全连接层

表一

用于音频设备标记的CNNS

VGGish [1]	CNN6	CNN10	CNN14
对数梅尔光谱仪 ram 96 frames × 64 mel bins	对数梅尔光谱仪 n 1000 frames × 64 mel bins		
3 3 64 ReLU	5 5 @ 64 BN, ReLU	(3 × 3 @ 64 BN, ReLU) × 2	(3 364 BN, ReLU) × 2
MP 2 × 2	Pooling 2 × 2		
3 3 128 ReLU	5 × 5 @ 128 BN, ReLU	(3 3 128 BN, ReLU) × 2	(3 3 128 BN, ReLU) × 2
MP 2 2	Pooling 2 × 2		
(3 × 3 @ 256 ReLU) 2.	5 × 5 @ 256 BN, ReLU	(3 × 3 @ 256 BN, ReLU) × 2	(3 3克256 BN, ReLU) × 2
MP 2 × 2	Pooling 2 × 2		
(3 × 3 @ 512 ReLU) × 2	5 5 512 BN, ReLU	(3 × 3 @ 512 BN, ReLU) × 2	(3 3 512 BN, ReLU) 2
MP 2 × 2 平坦 ，	全球汇集		泳池2 2
FC 4096 ReLU × 2	FC 512, ReLU		(3 3 @ 1024 BN, ReLU) × 2
FC 527, Sigmoid	FC 527, Sigmoid		Pooling 2 × 2
			(3 × 3 @ 2048 BN, ReLU) × 2
			全球的 pooling
			FC 2048, ReLU
			F527签名 noid

以进一步提高表现能力。我们研究了6-10层和14层CNN。基于AlexNet[34]，6层CNN由4个卷积层组成，核大小为5.5。10层和14层CNN分别由4层和6层卷积层组成，灵感来自类似VGG的CNN[35]。每个卷积块由2个卷积层组成，核大小为3.3。在每个卷积层之间应用批归一化[36]，并使用ReLU非线性[37]来加速和稳定训练。我们对每个卷积块应用大小为2 2的平均池进行下采样，因为2 2平均池已被证明优于2 2最大池[33]。

在最后一个卷积层之后应用全局池，将特征图汇总为固定长度的向量。在[15]中，全局池化使用了最大和平均操作。为了结合它们的优点，我们将平均向量和最大向量相加。在我们之前的工作[33]中，这些固定长度的向量被用作音频片段的嵌入特征。在这项工作中，我们在固定长度向量上添加了一个额外的全连接层来提取嵌入特征，这可以进一步提高它们的表示能力。对于特定的音频模式识别任务，线性分类器应用于嵌入特征，然后是用于分类任务的softmax非线性或用于标记任务的sigmoid非线性。在每次下采样操作和完全连接的层之后应用Dropout[38]，以防止系统过度拟合。表一总结了我们提出的CNN系统。“@符号后的数字表示特征图的数量。第一列显示了[13]提出的VGGish网络。MP是最大池化的缩写。表I中的“池化2 2”是大小为2 2的平均池化。在[13]中，一个音频片段被分割成1秒的片段，[13]还假设每个片段都继承了音频片段的标签，这可能会导致标签不正确。相比之下，我们的系统从表I中的第二列到第四列应用了整个音频片段进行训练，而没有将音频片段分割成片段。

表二

用于音频集标记的资源网

ResNet22	ResNet38	ResNet54
对数梅尔谱图1000帧64梅尔箱		
$(3 \times 3 @ 512, \mathrm{BN}, \mathrm{ReLU}) \times 2$		
池2 2		
基础B 64 2	基础B 64 3	瓶颈B 64 3
池2 2		
基础B 128 2	基础B1284	瓶颈B 128 4
池2 2		
基础B 256 2	基础B 256 6瓶装B 256 6	
池2 2		
基础B 512 2	基础B 512 3瓶装B 512 3	
池2	× 2	
（32048，BN，ReLU）2		
全球汇集		
FC 2048，ReLU		
FC 527，Sigmoid		

表三

用于音频集标记的移动网络

MobileNetV1	MobileNetV2
$3 \times 3 @ 32$, BN, ReLU	
池2 2	
V1区块@64 V1Block@128	V2块，t=1@16 （V2块，t=6@24）2
池2 2	
V1Block@128 V1区块@256	（V2块，t=6@32）3
池2 2	
V1区块@256 V1区块@512	（V2块，t=6@64）4
池2 2	
（V1区块@512）5 V1区块@1024	（V2块，t=6@96）3
池2 2	
V1区块@1024	V2区块，t=6160）3
	(V2Block, t=6 @ 320) × 1
全球汇集	
FC，1024，ReLU	
FC，527，Sigmoid	

我们将音频片段的波形表示为xn，其中n是音频片段的索引，f（n）[0,1]K是表示K个声音类别存在概率的PANN的输出。xn的标记表示为yn{0,1}K。使用二进制交叉熵损失函数l来训练PANNs：

$$l = -\overbrace{\sum_{n=1}^{N}}\ y_n \ln f_{xn} + 1 - y_n \ln 1 - f_{xn} \qquad 1.$$

其中N是AudioSet中的训练片段的数量。在训练中，通过使用梯度下降方法优化f（-）的参数，以最小化损失函数l。

B. 数据网

1） 传统残差网络（ResNets）：在音频分类方面，较深的CNN已被证明比较浅的CNN具有更好的性能[31]。非常深的传统卷积神经网络的一个挑战是，梯度不能从顶层正确地传播到底层[32]。为了解决这个问题，ResNets[32]在卷积层之间引入了快捷连接。这样，前向和后向信号可以直接从一层传播到任何其他层。快捷连接只引入了少量的额外参数和一点额外的计算复杂性。ResNet由几个块组成，其中每个块由两个卷积块组成。内核大小为3.3的层以及输入和输出之间的快捷连接。每个瓶颈块由三个卷积层组成，网络架构中有一个网络[39]，可以用来代替ResNet中的基本块[32]。

2） 调整ResNets进行AudioSet标记：我们将ResNet[32]调整为AudioSet标记，如下所示。首先，在log-mel频谱图上应用两个卷积层和一个下采样层，以减小输入log-mel谱图的大小。我们实现了三种不同深度的ResNet：一个22层的ResNet，有8个基本块；具有16个基本块的38层ResNet和具有16个残差块的54层ResNet。表II显示了适用于AudioSet标记的ResNet系统的架构。BasicB和BottleneckB分别是基本块和瓶颈块的缩写。

C. 移动网络

1） 传统移动网络：当系统在便携式设备上实现时，计算复杂性是一个重要问题。与CNN和ResNets相比，MobileNets旨在减少CNN中的参数数量和乘加操作，MobileNets基于深度可分离卷积[40]，通过将标准卷积分解为深度卷积和11点卷积[40]。

表三显示了AudioSet[40]的MobileNetV1和MobileNetV2[41]架构。V1Blocks和V2Blocks是MobileNet卷积块[40][41]，每个卷积块分别由两个和三个卷积层组成。

2) *Adapting MobileNets for AudioSet tagging* We adapt

D. 一维CNN

以前的音频标记系统基于log-mel频谱图，这是一种手工制作的功能。提高性能。一些研究人员提出构建直接对时域波形进行操作的一维CNN。例如，Dai等人[31]提出了用于声学场景分类的一维CNN，Lee等人[42]提出了一种一维CNN，后来Pons等人[15]将其用于音乐标签。

1） DaiNet:DaiNet[31]将长度为80、步幅为4的核应用于音频记录的输入波形。内核在训练过程中是可以学习的。首先，对第一卷积层应用最大值运算，该运算旨在使系统对输入信号的相移具有鲁棒性。然后，应用几个核大小为3、步幅为4的一维卷积块来提取高级特征。在UrbanSound8K分类中，每个卷积块中有四个卷积层的18层DaiNet取得了最佳结果[43][31]。

2） LeeNet：与在第一层应用大内核的DaiNet相反，LeeNet[42]在波形上应用了长度为3的小内核，以代替频谱图的STFT

提取。LeeNet由几个一维演化层组成，每个层后面都有一个大小为2的下采样层。原始的LeeNet由11层组成。

3） 将一维CNN用于AudioSet标记：
我们修改了LeeNet，将其扩展到具有24层的更深层次的架构，用由两个卷积层组成的卷积块替换每个卷积层。为了进一步增加一维CNN的层数，我们提出了一种核大小为3的一维残差网络（Res1dNet）。我们用残差块替换LeeNet中的卷积块，其中每个残差块由两个核大小为3的卷积层组成。卷积块的第一和第二卷积层分别具有1和2的膨胀，以增加相应残差块的感受野。在每个残差块之后应用下采样。通过使用14个和24个残差块，我们分别获得了31层和51层的Res1dNet31和Res1dNet 51

III、 波形-卷积神经网络系统

以前的一维CNN系统[31][42][15]并没有优于以log-mel谱图作为输入训练的系统。以前的时域CNN系统[31][42]的一个特点是，它们不是为了捕获频率信息而设计的，因为一维CNN系统中没有频率轴，所以它们无法捕获具有不同音调偏移的声音事件的频率模式。

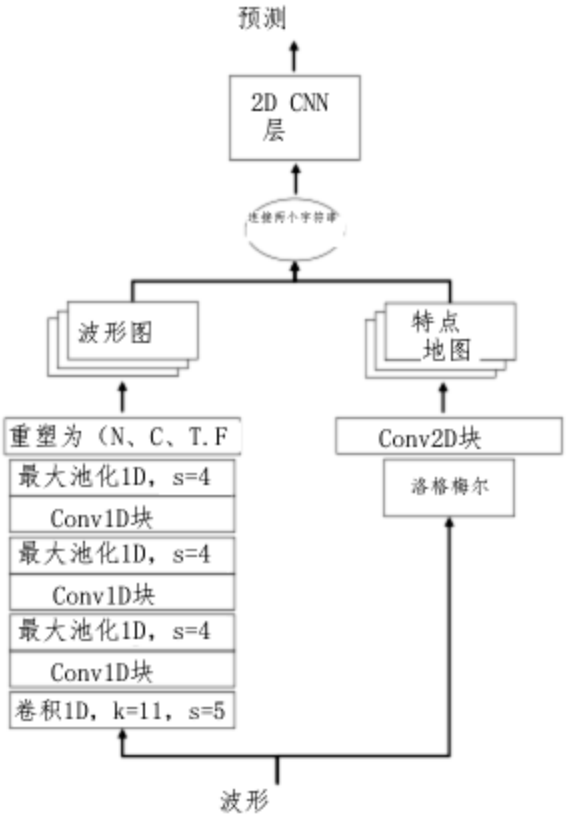

图1。Wavegram-Logmel CNN的架构

A. 波形-CNN系统

在本节中，我们提出了用于音频集标记的Wavegram CNN和Wavegram Logmel CNN架构。我们提出的Wavegram CNN是一种时域音频标记系统。我们提出的波形图是一种类似于log-mel频谱图的特征，但是使用神经网络学习的。波形图旨在学习傅里叶变换的修改后的时频表示。波形图具有时间轴和频率轴。频率模式对于音频模式识别很重要，例如，具有不同音调偏移的声音属于同一类。Wavegram旨在学习一维CNN系统中可能缺乏的频率信息。波图还可以通过从数据中学习一种新的时频变换来改进手工制作的log-mel谱图。然后，波形图可以代替对数梅尔谱图作为输入特征，从而形成我们的波形图-CNN系统。我们还将Wavegram和log-mel光谱图结合起来作为一个新特征，构建了Wavegram Logmel-CNN系统，如图1所示。

为了构建波形图，我们首先将一维CNN应用于时域波形。一维CNN从具有滤波器长度11和步长5的卷积层开始，以减小输入的大小。这会立即将输入长度减少5倍，以减少内存使用。接下来是三个卷积块，其中每个卷积块由两个卷积层组成，卷积层的膨胀分别为1和2，它们是设计用于增加卷积层的接收场。每个卷积块后面都有一个步幅为4的降采样层。通过使用步幅和降采样三次，我们将32kHz的音频记录降采样到每秒32000/5/4/4=100帧的特征。我们将一维CNN层的输出大小表示为TC，其中T是帧数，C是信道数。我们通过将C通道拆分为C/F组，将此输出重塑为T F C/F大小的张量，其中每组有F个频率区间。我们称这个张量为波图。Wavegram通过在每个C/F信道中引入F个频率仓来学习频率信息。我们在提取的Wavegram上应用第II-A节中描述的CNN14作为骨干架构，以便我们可以公平地比较基于Wavegram和log-mel频谱图的系统。二维CNN（如CNN14）可以捕获波形图上的时频不变模式，因为核在波形图中沿着时间和频率轴进行卷积。

B. Wavegram-Logmel-CNN

此外，我们可以将波形图和对数梅尔谱图组合成一个新的表示。通过这种方式，我们可以利用时域波形和对数梅尔谱图中的信息。组合是沿着通道维度进行的。Wavegram为音频标记提供了额外的信息，补充了log-mel频谱图。图1显示了Wavegram Logmel CNN的架构。

IV、 数据处理

在本节中，我们将介绍AudioSet标记的数据处理，包括数据平衡和数据增强。数据平衡是一种用于在高度不平衡的数据集上训练神经网络的技术。数据增强是一种用于增强数据集的技术，以防止系统在训练过程中过度拟合。

A. 数据平衡

可用于训练的音频片段数量因声音类别而异。例如，有90多万个音频片段属于“演讲”和“音乐”类别。另一方面，只有几十个音频片段属于“牙刷”类别。不同声音类别的音频片段数量呈长尾分布。在训练过程中，训练数据以小批量输入到PANN中。如果没有数据平衡策略，音频片段将从AudioSet中均匀采样。因此，在训练过程中更有可能对具有更多训练片段的声音类（如“语音”）进行采样。在极端情况下，一个小批量中的所有数据可能属于同一个声音类。这将导致PANN过度适应具有更多训练剪辑的声音类，而不足适应具有更少训练剪辑的音频类。为了解决这个问题，我们设计了一种平衡的采样策略来训练PANN。也就是说，从所有声音类中大致相等地采样音频剪辑以构成一个小批。我们使用“近似”一词是因为音频剪辑可能包含多个标签。

B. 数据扩充

数据增强是防止系统过度拟合的有效方法。AudioSet中的一些声音类只包含少量（例如数百个）训练片段，这可能会限制PANN的性能。我们在训练过程中应用mixup[44]和SpecAugment[45]来增强数据。

1） Mixup:Mixup[44]是一种通过插值数据集中两个音频片段的输入和目标来增强数据集的方法。例如，我们将两个音频片段的输入分别表示为x1，2，其目标分别表示为y1 y2。然后，可以分别通过z=Ax1+（1A）2和y=Ay1+（2A）y2获得增强输入和目标，其中A是从贝塔分布中采样的[44]。默认情况下，我们对log-mel频谱图应用mixup。我们将在第VI-C4节中比较log-mel频谱图和时域波形上的混频增强性能。

2） SpecAugment：SpecAugment[45]被提出用于对语音数据进行增强以进行语音识别。SpecAugment频率掩蔽和时间掩蔽，应用频率掩蔽，使得f个连续的梅尔频率区间[fo，fo+f]被掩蔽，其中f从0到频率掩蔽参数f的均匀分布中选择，fo从[0，Ff]中选择，其中f是梅尔频率区间的数量[45]。每个log-mel频谱图中可以有多个频率掩模。频率掩模可以提高PANN对音频片段频率失真的鲁棒性[45]。时间掩蔽类似于频率掩蔽，但应用于时域。。

operates on the log-mel spectrogram of an audio clip using

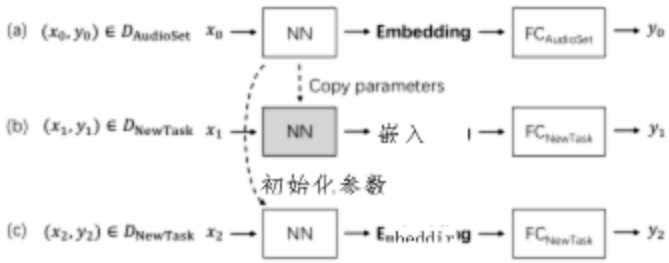

图2。（a） PANN使用AudioSet数据集进行预训练。（b） 对于新任务，PANN被用作特征提取器。基于提取的嵌入特征构建分类器。阴影矩形表示参数已冻结且未训练。（c） 对于新任务，神经网络的参数用PANN初始化。然后，对新任务的所有参数进行微调。

V. 转移到其他任务

为了研究PANNs的泛化能力，我们将PANNs转移到一系列音频模式识别任务中。之前关于音频迁移学习的研究[21][14][22][25][23]主要集中在音乐标记上，并且仅限于比AudioSet更小的数据集。首先，我们在图2（a）中演示了PANN的训练。这里，DAudioSet是AudioSet数据集，xo、yo分别是训练输入和目标。FCAudioset是AudioSet标签的全连接层。在这篇文章中，我们建议比较以下迁移学习策略。

1） 从头开始训练一个系统。所有参数都是随机初始化的。系统类似于PANN，除了最终的全连接层取决于任务相关的输出数量。该系统被用作与其他迁移学习系统进行比较的基线系统。

2） 使用PANN作为特征提取器。对于新任务
通过使用
潘。然后，嵌入特征被用作输入
分类器，如全连接神经网络
在训练新任务时，PANN的参数被冻结
没有受过训练。仅构建分类器的参数
对嵌入特征进行训练。图2（b）显示了这一点
策略，其中DNewTask是一个新的任务数据集，FCNewTask
是新任务的完全连接层。PANN用作
特征提取器。基于提取的嵌入构建分类器
丁特征。阴影矩形表示参数
它们是冷冻的，没有经过训练。。

（3） 微调PANN。PANN用于新任务，但最终的全连接层除外。所有参数都从PANN初始化，除了随机初始化的最终全连接层。所有参数都在DNewTask上进行了微调。图2（c）展示了PANN的微调。

VI、 实验

首先，我们评估了PANN在AudioSet标记上的性能。然后，将PANN转移到几个音频模式识别任务中，包括声学场景分类、通用音频标记、音乐分类和语音情感分类。

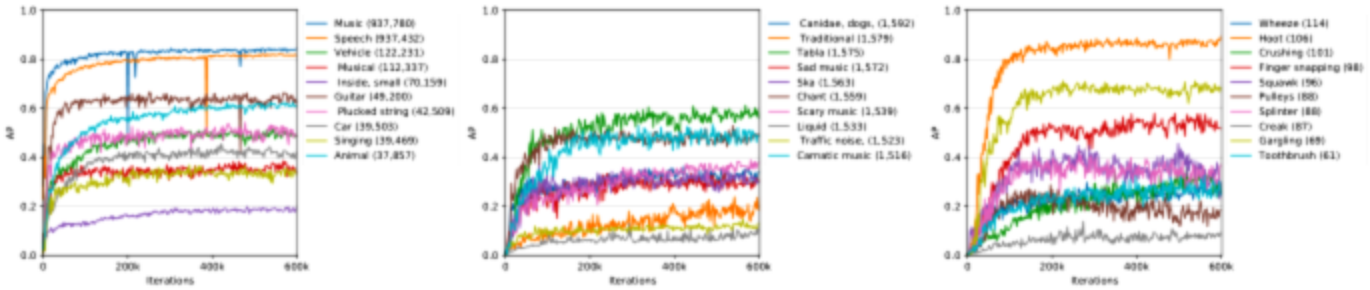

图3。使用CNN14系统对声音事件进行分级AP。括号内的数字表示训练片段的数量。左、中、右列显示了声音类的AP，训练片段的数量在AudioSet的训练集中排名第1至第10、第250至第260和第517至第527。

A. AudioSet数据集

AudioSet是一个大规模的音频数据集，包含527个声音类[1]。AudioSet中的音频片段是从YouTube视频中提取的。训练集由2063839个音频片段组成，包括22160个音频片段的平衡子集，其中每个声音类别至少有50个音频片段。评估集包含20371个音频片段。我们没有使用[1]提供的嵌入功能，而是在2018年12月通过[1]提供的链接下载了AudioSet的原始音频波形，并忽略了不再可下载的音频片段。我们成功下载了1934187个（94%）完整训练集的音频片段，包括20550个（93%）平衡训练集的视频片段。我们成功下载了18887个评估数据集的音频片段。如果音频片段短于10秒，我们会将其静音至10秒。考虑到YouTube上的大量音频片段是单声道的，采样率低，我们将所有音频片段转换为单声道，并将其重新采样为32 kHz。

对于基于log-mel谱图的CNN系统，STFT应用于大小为1024的汉明窗口[33]和跳数为320个样本的波形。这种配置导致每秒100帧。根据[33]，我们应用64个梅尔滤波器组来计算对数梅尔谱图。梅尔组的下限和上限截止频率分别设置为50 Hz和14 kHz，以消除低频噪声和混叠效应。我们使用torchlibrosa，librosa[46]函数的PyTorch实现，将log-mel光谱图提取构建到PANNs中。10秒音频片段的log-mel频谱图具有1001 64的形状。额外的一帧是在计算STFT时应用中心参数造成的。批量大小为32，使用学习率为0.o01的Adam[47]优化器进行训练。系统在一张Tesla -V100-PCI-32GB卡上进行训练。每个系统需要大约3天的时间从头开始训练600k次迭代。

B. 评价指标

平均精度（mAP）、平均曲线下面积（mAUC）和d-prime被用作AudioSet标记的官方评估指标[20][1]。平均精度（AP）是

表四
与以往方法的比较

	mAP	AUC	d-prime
随机猜测	0.005	0.500	0
谷歌CNN[1]	0.314	0.959	2.452
单级关注[16]	0.337	0.968	2.612
多层次关注[17]	0.360	0.970	2.660
功能级别关注度高[20]	0.369	0.969	2.640
好未来网[19]	0.362	0.965	2.554
DeepRes[48]	0.392	160	2.682
我们提出的CNN14	0.431	0.973	2.732

召回率和精确度曲线下的面积。AP不依赖于真阴性的数量，因为精确性和召回率都不依赖于真实阴性的数量。另一方面，AUC是假阳性率和真阳性率（召回率）下的面积，反映了真阴性的影响。d-prime[1]也用作度量，可以直接从AUC[1]计算。所有指标都是在单个类上计算的，然后在所有类上取平均值。这些指标也称为宏观指标。

C. AudioSet标记结果

I） 与先前方法的比较：表四显示了我们提出的CNN14系统与先前Au-dioSet标记系统的比较。我们选择CNN14作为基本模型来研究AudioSet标记的各种超参数配置，因为CNN14是一个结构简单的标准CNN，可以与之前的CNN系统进行比较[3][33]。作为基线，随机猜测分别达到0.005的mAP、0.500的AUC和0.000的d-prime。谷歌[1]发布的结果使用[13]中的嵌入特征进行训练，分别达到了0.314的mAP和0.959的AUC。单级注意力和多级注意力系统[16]、[17]实现了0.337和0.360的mAP，后来通过特征级注意力神经网络进行了改进，实现了0.369的mAP。Wang等人[19]研究了五种不同类型的注意力功能，并获得了0.362的mAP。上述所有系统都是基于AudioSet发布的嵌入功能构建的[1]。最近的DeepRes系统[48]是基于从YouTube下载的波形构建的，并实现了0.392的mAP。表四的底部行显示了我们提出的

*https://github.com/qiuqiangkong/torchlibrosa

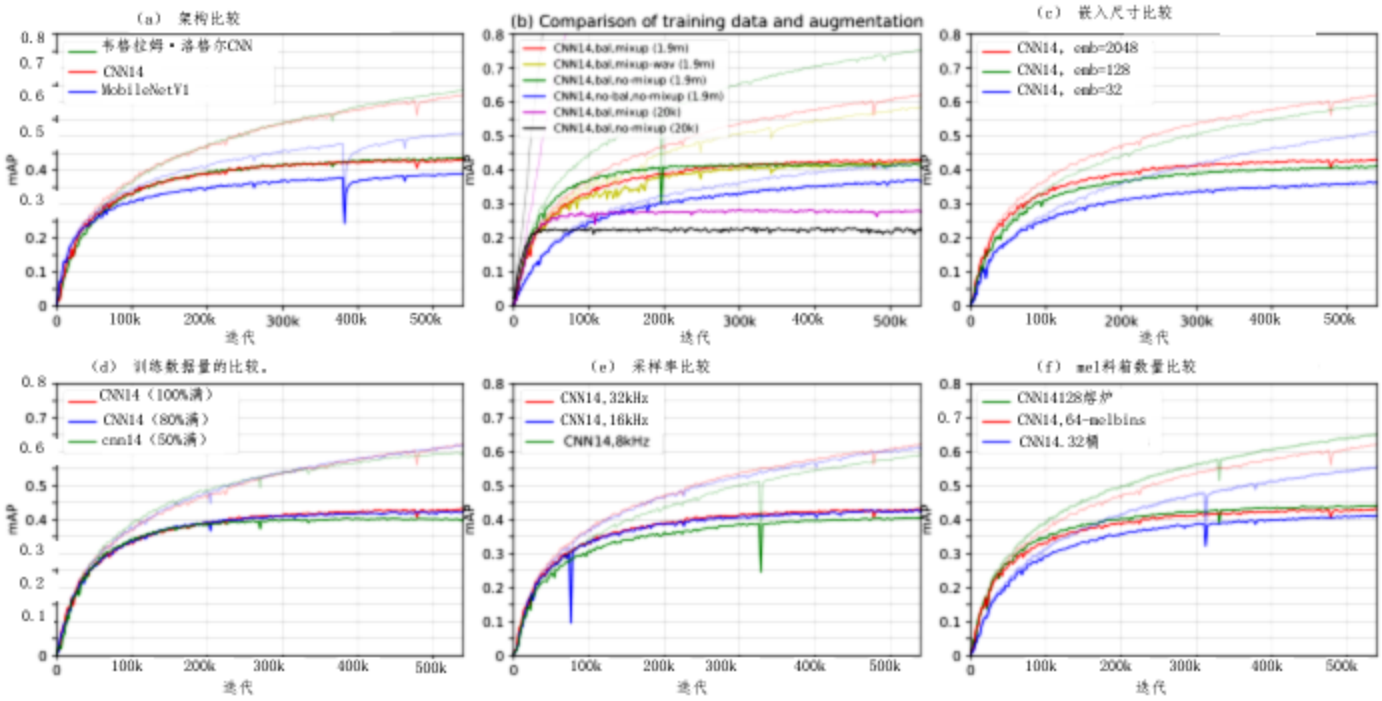

图4.音频集标记的PANNs结果。透明和实线分别是训练mAP和评估mAP。六个图显示了不同的结果：（a）架构；（b）数据平衡和数据增强；（c）嵌入尺寸；（d）训练数据量；（e）采样率；（f）梅尔垃圾桶的数量

CNN14系统的mAP为0.431，优于之前最好的系统。我们使用CNN14作为骨干来构建Wavegram-Logmel CNN，以便与CNN14系统进行公平比较。图4（a）显示，Wavegram-Logmel CNN的性能优于CNN14系统和MobileNetV1系统。详细结果见本节后面的表XI

2） 按类别性能：图3显示了CNN14系统不同声音类别的按类别AP，左、中、右列显示了声音类的AP，其中训练片段的数量在AudioSet的训练集中排名第1至第1位，第250至第260位和第517至第527位。不同声音等级的表现可能非常不同。例如，“音乐”和“语音”的AP超过0.80。另一方面，一些声音类，如“Inside，small”，只能达到0.19的AP。图3显示，AP通常与训练片段的数量无关。例如，左列显示“Inside，small”包含70159个训练片段，而其AP较低。相比之下，右栏显示“Hoot只有106个训练片段，但达到了0.86的AP，并且比许多其他训练片段更多的声音类别更大。在本文的最后，我们在图12中绘制了所有527个声音类别的mAP，图12显示了CNN14、MobileNetV1和Wavegram Logmel CNN系统与之前最先进的音频标记系统[20]的逐类比较，该系统使用[1]发布的嵌入功能构建。图12中的蓝色条显示了训练片段的对数数量，“+”符号表示0和1之间的标记质量，这是通过专家验证的正确标记的音频片段的百分比来衡量的[1]。标签质量因声音等级而异，“_”符号表示

表五

数据平衡和增强的结果

增强	mAP	AUC	d-prime
无平衡，无混淆（20k）	0.224	0.894	1.763
bal，无混音（20k）	0.221	0.879	1.652
混球（20k）	0.278	0.905	1.850
无平衡，无混淆（1.9米）	0.375	0.971	2.690
平衡，无混淆（1.9米）	0.416	0.968	2.613
混合球（1.9米）	0.431	0.973	2.732
混球wav（1.9米）	0.425	0.973	2.720

表VI

不同啤酒花大小的结果

跳跃大小	时间分辨率	mAP	AUC	d-prime
1000	31.25毫秒	0.400	0.969	2.645
640	20.00毫秒	0.417	0.972	2.711
500	15.63毫秒	0.417	0.971	2.682
320	10.00毫秒	0.431	0.973	2.732

标签质量不可用的声音类别。图12显示，某些类别的平均精度高于其他类别。例如，“风笛”等声音类别的平均精度为0.90，而“屁股”等声音等级的平均精度小于0.2。一种解释是，不同声音类别的音频标记难度不同。此外，音频标记性能并不总是与训练片段的数量和标签质量相关[20]。图12显示，我们提出的系统在各种声音类别上都优于之前最先进的系统[16]、[17]。

表七
不同嵌入维度的结果

嵌入	mAP	AUC	d-prime
32	0.364	0.958	2.437
128	0.412	0.969	2.634
512	0.420	160	2.689
2048	0.431	0.973	2.732

表八
部分训练数据的结果

培训数据	mAP	AUC	d-prime
满的50%	0.406	0.964	2.543
80%满	0.426	0.971	2.677
100%满	0.431	0.973	2.732

3） 数据平衡：第IV-A节介绍了我们用于训练AudioSet标记系统的数据平衡技术。图4（b）显示了有和没有数据平衡的CNN14系统的性能。蓝色曲线表明，在没有数据平衡的情况下训练PANN需要很长时间。绿色曲线表明，通过数据平衡，系统在有限的训练迭代内收敛得更快。此外，用完整的190万个训练片段训练的系统比用20k个训练片段的平衡子集训练的系统表现更好。表五显示，CNN14系统在数据平衡的情况下实现了0.416的mAP，高于没有数据平衡的mAP（0.375）。

4） 数据增强：我们发现混合数据增强在训练PANNs中起着重要作用。默认情况下，我们对log-mel频谱图应用mixup。图4（b）和表V显示，用混淆数据增强训练的CNN14系统达到了0.431的mAP，优于没有混淆数据增强的训练系统（0.416）。当使用仅包含20k个训练片段的平衡子集进行训练时，与没有混淆的训练（0.221）相比，混淆特别有用，产生0.278的mAP。此外，我们还表明，当使用完整的训练数据进行训练时，log-mel谱图上的混音达到了0.431的mAP，优于0.425时域波形中的混音。这表明，当与log-mel频谱图一起使用时，混合比与时域波形一起使用时更有效。5） 跳跃大小：跳跃大小是样本的数量

在相邻帧之间。跳数越小，跳数越高
时域分辨率。我们调查影响
使用CNN14在AudioSet标签上标记不同的跳数
系统。我们研究了1000、640、500和320的跳数大小：
这些对应于31.25ms的时域分辨率，
相邻帧之间的间隔为20.00ms、15.63ms和10.00ms，
分别。表VI显示，mAP评分随着
跳跃大小减小。CNN14系统的跳数为320
达到0.431的mAP，优于较大的跳数
例如500、640和1000。

6） 嵌入尺寸：嵌入特征是固定的-
总结音频片段的长度向量。默认情况下
CNN14的嵌入特征维度为2048。我们
研究一系列具有嵌入二聚体的CNN14系统-

表IX
不同采样率的结果

采样率	mAP	AUC	d-prime
8千赫	0.406	0.970	2.654
16千赫	0.427	0.973	2.719
32千赫	0.431	0.973	2.732

表X
不同MEL箱的结果

梅尔·宾斯	mAP	AUC	d-prime
32个箱子	0.413	0.971	2.691
64个箱子	0.431	0.973	2.732
128个箱子	0.442	0.973	2.735

32、128、512和2048。图4（c）和表VII显示，mAP性能随着嵌入尺寸的增加而增加。

7） 部分数据训练：AudioSet的音频片段来自YouTube。相同的音频片段不再可下载，其他片段将来可能会被删除。为了在未来更好地再现我们的工作，我们研究了用随机选择的部分数据（从下载数据的50%到100%）训练的系统的性能。图4（d）和表VIII显示，当CNN14系统用80%的完整数据进行训练时，mAP从0.431略微下降到0.426（下降1.2%），当用50%的完整数据训练时，下降到0.406（下降5.8%）。这一结果表明，训练数据量对训练PANN很重要。

8） 采样率：图4（e）和表IX显示了用不同采样率训练的CNN14系统的性能。用16 kHz录音训练的CNN14系统达到0.427的mAP，与用32 kHz采样率训练的CNN114系统相似（在1.0%以内）。另一方面，用8kHz录音训练的CNN14系统实现了0.406的较低mAP（降低5.8%）。这表明4kHz-8kHz范围内的信息对于音频标记很有用。

9） 梅尔仓：图4（f）和表X显示了用不同数量的梅尔仓训练的CNN14系统的性能。该系统在32个梅尔箱中实现了0.413的mAP，而在64个梅尔箱和128个梅尔箱的情况下分别为0.431和0.442。这一结果表明，尽管计算复杂度随梅尔箱的数量呈线性增加，但PANN在梅尔箱越多的情况下性能越好。在本文中，我们采用64个梅尔箱来提取对数梅尔谱图，作为计算复杂度和系统性能之间的权衡。

10） CNN层数：如第II-A节所述，我们研究了具有6层、10层和14层的CNN系统的性能。表XI显示，6、10和14层CNN分别实现了0.343、0.380和0.431的mAP。这一结果表明，具有较深CNN架构的PANN比较浅CNN架构的性能更好。这一结果与之前在较小数据集上训练的音频标记系统形成鲜明对比

表XI
不同系统的结果

建筑	mAP	AUC	d-prime
CNN6	0.343	0.965	2.568
CNN10	0.380	0.971	2.678
CNN14	0.431	0.973	2.732
ResNet22	0.430	0.973	0.270
ResNet38	0.434	0.974	2.737
ResNet54	0.429	0.971	2.675
MobileNetV1	0.389	0.970	2.653
MobileNetV2	0.383	0.968	2.624
DaiNet[31]	0.295	0.958	2.437
LeeNet11[42]	0.266	0.953	2.371
LeeNet24	0.336	0.963	2.525
Res1dNet31	0.365	0.958	2.444
Res1dNet51	0.355	0.948	2.295
美国有线电视新闻网Wavegram	0.389	0.968	2.612
韦格拉姆·洛格尔CNN	0.439	0.973	2.720

表十二
不同系统的多重加法次数和参数

建筑	多添加	参数
CNN6	二十一点九八六一零	4,837,455
CNN10	二十八点一六六一零九	5,219,279
CNN14	四十二点二二零一零	80,753,615
ResNet22	三十点零八一一零九	63,675,087
ResNet38	四十八点九六二一零九	73,783,247
ResNet54	五十四点五六三一零九	104,318,159
MobileNetV1	三点六一四一零九	4,796,303
MobileNetV2	二点八一零一零	4,075,343
DaiNet	三十点三九五一零九	4,385,807
LeeNet11	4.741×10^9	748,367
LeeNet24	二十六点三六九一零九	10,003,791
Res1dNet31	三十二点六八八一零九	80,464,463
Res1dNet51	六十一点八三三一零九	106,538,063
美国有线电视新闻网Wavegram	四十四点二三四一零九	80,991,759
韦格拉姆·洛格尔CNN	五十三点五一零一零九	81,065,487

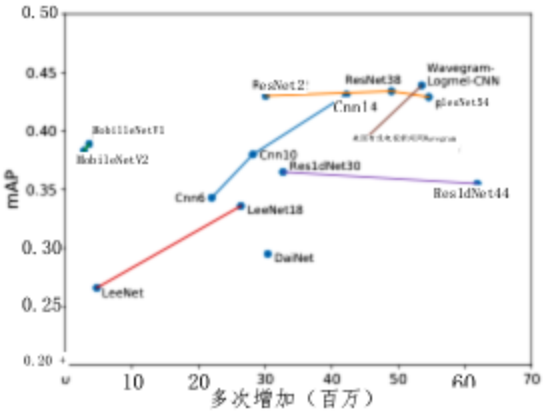

图5。AudioSet标签系统的多重添加与mAP。相同类型的架构以相同的颜色分组

9层CNN等CNN的表现优于深层CNN[33]。一种可能的解释是，较小的数据集可能会受到过拟合的影响，而AudioSet足够大，可以训练更深层次的CNN，至少可以训练到我们研究的14层CNN。

11） ResNets：我们应用ResNets来研究更深层次PANN的性能。表XI显示，ResNet22系统实现了与CNN14系统类似的0.430的mAP。ResNet38的mAP为0.434，略高于其他系统。ResNet54实现了0.429的mAP，这并没有进一步提高性能。。

12） MobileNets：上述系统表明，PANNs在AudioSet标记方面取得了良好的性能。然而，当在便携式设备上实现时，这些系统没有考虑计算效率。我们研究了第II-C节中描述的使用轻量级MobileNets构建PANN。表XI显示MobileNetV1实现了0.389的mAP，比CNN14系统0.431低9.7%。MobileNetV1系统的乘法和加法（多重加法）数量和参数分别仅为CNN14系统的8.6%和5.9%。MobileNetV2系统实现了0.383的mAP，比CNN14低11.1%，并且比MobileNetV1的计算效率更高，其中多重添加和参数的数量仅为CNN14系统的6.7%和5.0%。

13） 一维CNNs：表XI显示了一维CNNs的性能。具有18层的DaiNet[31]实现了0.295的mAP。具有11层的LeeNet11[42]实现了0.266的mAP。我们改进的24层LeeNet将LeeNet11的mAP提高到0.336。我们在第II-D3节中提出的Res1dNet31和Res1dNet 51分别实现了0.365和0.355的mAP，并在一维CNN系统中实现了最先进的性能。

14） Wavegram-Logmel-CNN：表XI的底行显示了我们提出的Wavegram-CNN和Wavegral-Logmel-CNN系统的结果。Wavegram CNN系统实现了0.389的mAP，优于之前最好的一维CNN系统（Res1dNet31）。这个表明波形图是一种有效的学习特征。此外，我们提出的Wavegram-Logmel CNN系统在所有PANN中实现了0.439的最先进mAP。

15） 复杂性分析：我们分析了用于推理的PANN的计算复杂性。多重加法的数量和参数是复杂性分析的两个重要因素。表XII的中间列显示了推断10秒音频片段的多重相加次数。表XII的右栏显示了不同系统的参数数量。CNN14系统的多加数和参数分别为42210万和8080万，大于CNN6和CNN10系统。ResNets22和ResNet38系统的多重添加次数略少于CNN14系统。ResNet54系统包含的多重加法最多，为54.6 10。一维CNN的计算成本与二维CNN相似。性能最好的一维系统Res1dNet31包含32.7 10个多重加法和8050万个参数。我们提出的Wavegram CNN系统包含44.2 10个多加法和8100万个参数，与CNN14相似。Wavegram-Logmel CNN系统将倍数略微增加到53.5 10，参数数量为8110万，与CNN14相似。为了减少多重加法和参数的数量，

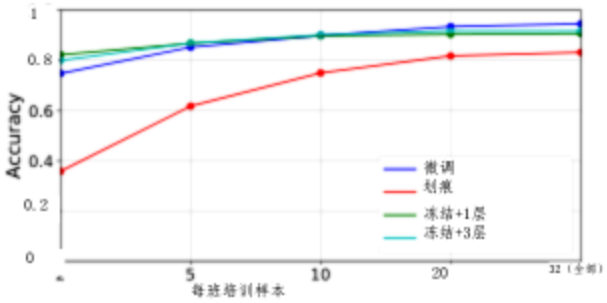

图6。ESC-50的准确性，每节课有不同数量的训练片段。

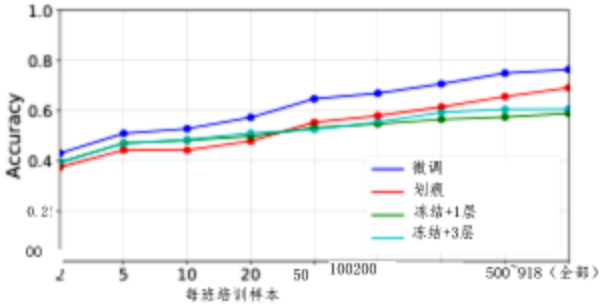

图7:DCASE 2019任务1的准确性，每节课有不同数量的训练片段。

表十三
ESC-50的精度

	斯托阿[49]	划痕	微调	免费_L1	Freeze_L3
Acc。	0.865	0.833	0.947	0.908	0.918

应用了移动网络，MobileNetV1和MobileNetV2系统是轻量级的卷积神经网络，分别只有3.6 109和2.810个多重加法和约480万和410万个参数。移动网络降低了计算成本和系统规模。图5总结了不同PANN的mAP与多次添加的mAP。相同类型的系统由相同颜色的线条连接。mAP从下到上增加。右上角是我们提出的Wavegram-Logmel CNN系统，它实现了最佳的mAP。左上角是计算效率最高的系统MobileNetV1和MobileNetV2。

表十四
DCASE 2019任务1的准确性

	斯托阿[51]	划痕	微调	免费_L1	Freeze_L3
Acc。	0.851	0.691	0.764	0.589	0.607

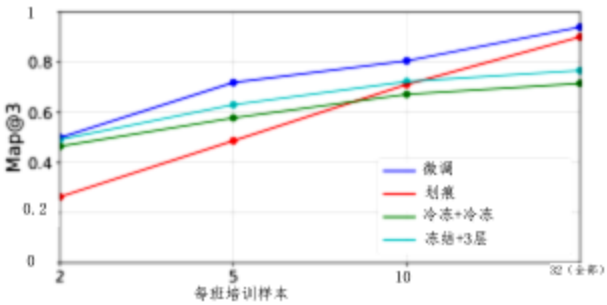

图8.DCASE 2018任务2的准确性，每节课有不同数量的训练片段。

D.转到其他任务

在本节中，我们将研究PANN在一系列其他模式识别任务中的应用。对于只提供有限数量训练片段的任务，PANN可用于少镜头学习。很少有镜头学习是音频模式识别中的一个重要研究课题，因为收集标记数据可能很耗时。我们使用第五节中描述的方法将PANN转移到其他音频模式识别任务中。首先，我们将所有音频记录重新采样到32 kHz，并将其转换为单声道，以与在AudioSet上训练的PANN保持一致。我们为每项任务执行第五节中描述的以下策略：1）从头开始训练系统；2） 使用PANN作为特征提取器；3） 微调PANN。当使用PANN作为特征提取器时，我们在具有一个和三个完全连接层的嵌入特征上构建分类器，分别称为Freeze_L1和Freeze_L_3。我们采用CNN14系统进行迁移学习，以便与其他基于CNN的音频模式识别系统进行公平的比较。我们还研究了在训练其他音频模式识别任务时，用不同数量的镜头训练的PANN的性能。

表XV
DCASE 2018任务2的准确性

	斯托阿[52]	划痕	微调	免费_L1	Freeze_L3
mAP@3	0.954	0.902	0.941	0.717	0.768

每节课40个片段。表11显示了CNN14系统的5倍交叉验证[50]精度。Sailor等人[49]提出了一种最先进的ESC-50系统，使用卷积受限玻尔兹曼机进行无监督滤波器组学习，精度达到0.865。我们的微调系统实现了0.947的精度，大大优于之前最先进的系统。Freeze_L1和Freeze_L3系统分别实现了0.918和0.908的精度。从头开始训练CNN14系统达到了0.833的精度。图6显示了ESC-50在每个声音类别的不同训练片段数量下的精度。当每个声音类别可用于训练的片段少于10个时，使用PANN作为特征提取器可以获得最佳性能。通过更多的训练片段，微调后的系统可以实现更好的性能。微调系统和使用PANN作为特征提取器的系统都优于从头开始训练的系统。

1） ESC-50：ESC-50是一个环境声音数据集[50]，由50个声音事件组成，如“狗”和“雨”，数据集中有2000个5秒的音频片段

2） DCASE 2019任务1：DCASE 2019的任务1是一个声学场景分类任务[2]，其数据集由以下部分组成

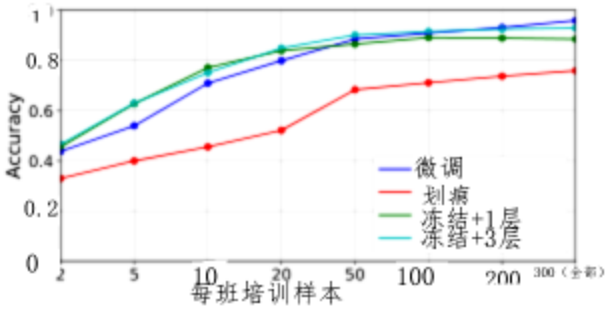

图9。MSoS的准确性，每节课有不同数量的训练片段。

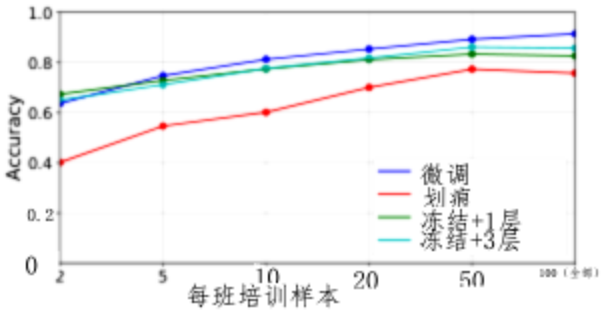

图10。GTZAN的准确性，每节课有不同数量的训练片段

表XVI
MSOS的准确性

	斯托阿[53]	划痕	微调	免费 L1	Freeze_L3
Acc。	0.930	0.760	0.960	0.886	0.930

表十七
GTZAN的准确性

	斯托阿[56]	划痕	微调	免费 L1	Freeze_L3
Acc。	0.939	0.758	0.915	0.827	0.858

从12个欧洲城市的各种声学场景中收集了40小时的立体声录音。我们专注于子任务A，其中每个音频记录都有两个通道，采样率为48kHz。在开发集中，分别有9185个和4185个音频片段用于训练和验证。我们通过平均立体声声道将立体声录音转换为单声道。从头开始训练的CNN14达到了0.691的精度，而微调系统达到了0.764的精度。Freeze_L1和Freeze_L3分别实现了0.689和0.607的精度，并且没有超过从头开始训练的CNN14。这种表现不佳的一种可能解释是，声学场景分类的录音具有不同的AudioSet分布。因此，使用PANN作为特征提取器并不能比从头开始微调或训练系统更好。微调后的系统比从头开始训练的系统性能更好。图7显示了每类具有不同数量训练片段的系统的分类精度。表十四显示了总体业绩。Chen等人[51]的最先进系统。使用各种分类器和立体声录音的组合作为输入，实现了0.851的精度，而我们不使用任何集成方法和立体声录音。

3） DCASE 2018任务2:DCASE 2018任务2是一个通用的自动音频标记任务[54]，使用Freesound的录音数据集，并用AudioSet本体中的41个标签的词汇表进行注释。开发集由9473个录音组成，持续时间从300毫秒到30秒mAP@3用于评估系统性能[54]。在训练中，我们将音频记录分成4秒的音频片段。在推理中，我们对这些片段的预测进行平均，以获得的预测。录音。表十五显示，郑和林[52]提出的最佳先前方法实现了mAP@30.954使用了几个系统的集成。相比之下，我们从头开始训练的CNN14系统达到了0.902的精度。经过微调的CNN14系统实现了mAP@30.941。Freeze_L1和Freeze_L3系统分别实现了0.717和0.768的精度。图8显示了

mAP@3使用不同数量的训练片段。微调后的CNN14系统优于从头开始训练的系统和使用PANN作为特征提取器的系统。微调的CNN14系统实现了与最先进系统相当的结果。

4） MSoS:声音的意义（MSoS）数据挑战[55]是一项将录音预测为五类之一的任务：“自然”、“音乐”、“人类”、效果和城市。该数据集由1500个音频片段的开发集和500个音频片段组成的评估集组成。所有音频片段的持续时间为4秒。陈和古普塔[53]提出的最先进的系统达到了0.930的精度。我们经过微调的CNN14达到了0.960的精度，优于之前最先进的系统。从头开始训练的CNN14达到了0.760的精度。图9显示了不同训练片段数量的系统的准确性。微调的CNN14系统和使用CNN14作为特征提取器的系统优于从头开始训练的CNN14。

5） GTZAN:GTZAN数据集[57]是一个音乐流派分类数据集，包含10种音乐流派的1000个30秒音乐片段，如古典音乐和乡村音乐。所有音乐片段的持续时间为30秒，采样率为22050 Hz。在开发过程中，使用10倍交叉验证来评估系统的性能。表XVII显示，Liu等人[56]提出的先前最先进的系统使用自下而上的广播神经网络实现了0.939的精度。微调的CNN14系统实现了0.915的精度，优于从头开始训练的CNN1四，精度为0.758，Freeze_L1和Freeze_L3系统，精度分别为0.827和0.858。图10显示了具有不同数量训练片段的系统的准确性。Freeze_L1和Freeze_L3系统的性能优于其他每班训练少于10个剪辑的系统。通过更多的训练片段，微调后的CNN14系统比其他系统表现更好。

6） RAVDESS：瑞尔森情感语音和歌曲视听数据库（RAVDESS）是一个人类语音情感数据集[59]。数据库由以下声音组成

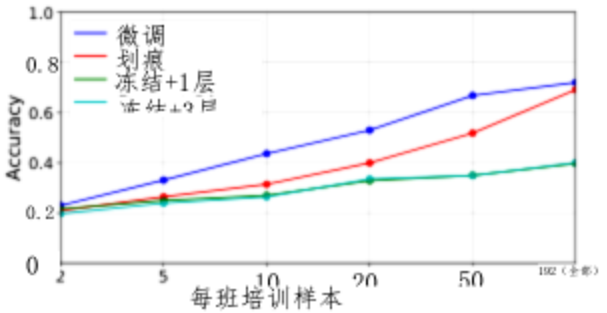

图11。每节课有不同数量的训练片段，RAVDESS的准确性。

表十八
雷达的准确性

	斯托阿[58]	划痕	微调	免费 L1	Freeze_L3
Acc。	0.645	0.692	0.721	0.397	0.401

24名专业演员，包括12名女性和12名男性，模拟8种情绪，如快乐和悲伤。任务是将每个声音片段分类为一种情感。开发集中有1440个音频片段。我们通过4倍交叉验证来评估我们的系统。表XVIII显示，曾等人[58]提出的先前最先进的系统达到了0.645的精度。我们从头开始训练的CNN14系统达到了0.692的精度。微调的CNN14系统达到了0.721的最先进精度。Freeze_L1和Freeze_L3系统分别达到了0.397和0.401的较低精度。图11显示了系统相对于一系列训练片段的准确性。微调的系统和从头开始训练的系统。其性能优于使用PANN作为特征提取器的系统。这表明RAVDESS数据集的录音可能具有不同的AudioSet数据集分布。因此，需要对PANN的参数进行微调，以在RAVDESS分类任务中实现良好的性能。

E. 讨论

在这篇文章中，我们研究了用于AudioSet标记的各种PANN。我们提出的几个PANN的表现优于之前最先进的AudioSet标记系统，包括CNN14的mAP达到0.431，ResNet38的mAP为0.434，优于谷歌的0.314基线。MobileNets是轻量级的系统，具有较少的多重加法和参数数量。MobileNetV1的mAP达到了0.389。我们改进的一维系统Res1dNet31的mAP为0.365，优于之前的一维CNN，包括0.295的DaiNet[31]和0.266的LeeNet11[42]。我们提出的Wavegram-Logmel CNN系统在所有PANN中实现了最高的0.439 mAP。PANNs可以用作新音频模式识别任务的预训练模型。

在AudioSet数据集上训练的PANN被转移到六个音频模式识别任务中。我们证明，经过微调的PANN在ESC-50、MSOS和RAVDESS分类任务中实现了最先进的性能，并接近在DCASE 2018任务2和GTZAN分类任务中取得了最先进的性能。在PANN系统中，经过微调的PANN在新任务上的表现总是优于从头开始训练的PANN。实验表明，PANNs在有限训练数据的情况下成功地推广到其他音频模式识别任务。。

七、结论

我们提出了在AudioSet上训练的预训练音频神经网络（PANNs），用于音频模式识别。人们研究了各种神经网络来构建PANN。我们提出了一种从波形中学习的波形图特征，以及一种在AudioSet标记中实现最先进性能的波形图Logmel CNN，存档了0.439的mAP。我们还研究了PANN的计算复杂性。我们证明，PANN可以转移到广泛的音频模式识别任务中，并优于之前几种最先进的系统。当对新任务的少量数据进行微调时，PANN可能很有用。未来，我们将把PANN扩展到更多的音频模式识别任务。

参考文献

[1] J.F.Gemmeke，D.P.Ellis，D.Freedman，A.Jansen，W.Lawrence，R.C。Moore，M.Plakal和M.Ritter，*音频集：本体论和人类-标记音频事件的标记数据集”，IEEE国际会议声学、语音和信号处理（ICASSP），2017，第776780页。

[2] A.Mesaros、T.Heittola和T.Virtanen，一个多设备数据集城市声学场景分类”，在探测和声学场景和事件分类（DCASE），2018，第9页13.

[3] K.Choi、G.Fazekas和M.Sandler，“使用深度自动标记卷积神经网络，”在国际会议音乐信息检索学会（ISMIR），2016，第805811页。

[4] E.Cakir、T.Heittola、H.Huttunen和T.Virtanen，复调声音国际上使用多标签深度神经网络的事件检测神经网络联合会议（IJCNN），2015年。

[5] J.P.Woodard，按产品分类的自然声音建模和分类代码隐马尔可夫模型、IEEE信号处理学报，1992年，第40卷，第18331835页。

[6] D.P.W.Ellis，检测警报声，https://academiccommons.哥伦比亚.edu/doi/10.7916/D8F19821/下载，2001年。

[7] D.斯托维尔、D.吉安努利斯、E.贝内托斯、M.拉格朗日和M.D。Plumbley，声学场景和事件的检测和分类，IEEE多媒体汇刊，第17卷，第173317462015页。

[8] A.梅萨罗斯、T.海托拉、E.贝内托斯、P.福斯特、M.拉格朗日、T.维尔塔宁。和M.D.Plumhlev，*声学场景的检测和分类和活动：2016年DCASE挑战赛的结果，IEEE/ACM音频、语音和语言处理学报（TASLP），第26卷，第379393页，2018年。

[9] A.梅萨罗斯、T.海托拉、A.迪蒙特、B.伊丽莎白、A.沙阿、E.文森特，B.Raj和T.Virtanen，DCASE 2017挑战设置：任务、数据集以及基线系统”，在关于检测和分类的研讨会上声学场景与事件（DCASE），2017，第8592页。

[10] 2019年DCASE挑战赛，http://dcase.community/challenge2019，2019.

[11] 邓、董、索彻、李、李、费飞，ImageNet：IEEE大会上的大规模分层图像数据库计算机视觉与模式识别（CVPR），2009，第248255页。

[12] J.Devlin，M.-W.Chang，K.Lee和K.Toutanova，BERT：预-训练用于语言理解的深层双向转换器，协会北美分会年会计算语言学（NAACL），2018，第41714186页。

[13] S.好时、S.乔杜里、D.P.埃利斯、J.F.杰梅克、A.詹森、R.C。Moore、M.Plakal、D.Platt、R.A.Saurous、B.Seybold等人，美国有线电视新闻网IEEE国际标准中的“大规模音频分类架构”2017年声学、语音和信号处理会议（ICASSP），第131135页。

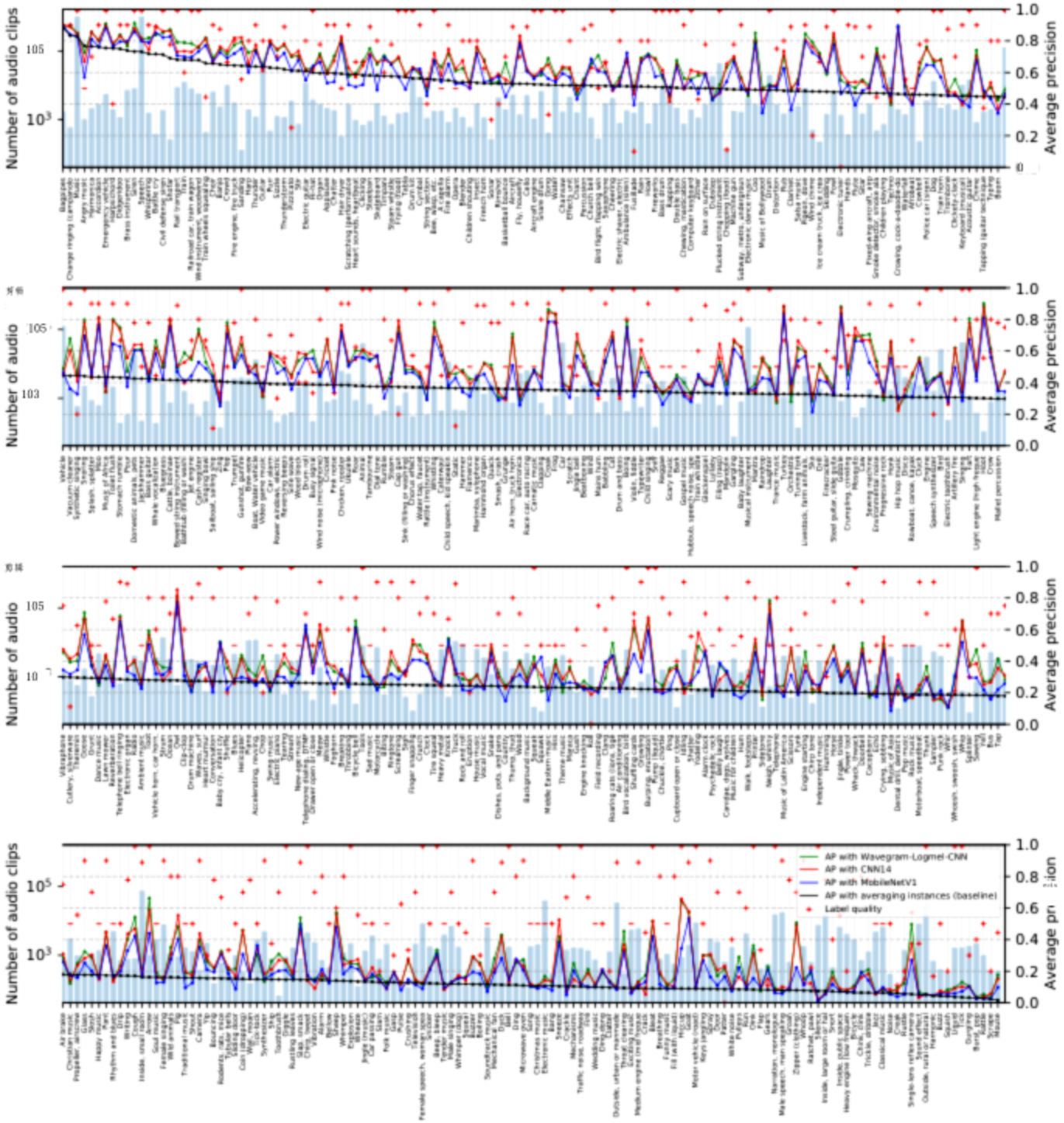

图12。AudioSet标签系统的分类性能。红色、蓝色和黑色曲线是CNN14、MobileNetV1和音频标记系统的AP[20]。蓝色条以对数刻度显示训练片段的数量。

[14] K. Choi、G. Fazekas、M. Sandler和K. Cho，迁移学习
音乐分类和回归任务”，在会议
国际音乐信息检索学会（ISMIR），2017，pp。
141149.

[15] J. Pons、O. Nieto、M. Prockup、E. Schmidt、A. Ehmann和X. Serra
在Conference中大规模进行音乐音频标签的端到端学习
国际音乐信息检索学会（ISMIR），
2017年，第637644页。

[16] 孔、徐、王和普拉布利，“音频设备分类”
《注意力模型：概率视角》，发表在IEEE International上
声学、语音和信号处理会议（ICASSP），2018年，
第316320页。

[17] C. Yu，K. S. Barsim。孔和杨。多层次关注
弱监督音频分类模型，”在检测和
声学场景和事件分类（DCASE），2018，第188页
192.

[18] 张，杨。学会识别
使用注意力监督的瞬态声音事件，”International，
人工智能联合会议（IJCAI），2018，第3336页
3342.

[19] 王、李、梅泽：五个多实例的比较
学习用于弱标签声音事件检测的池函数，”

IEEE声学、语音和信号国际会议
《加工》（ICASSP），2019年，第3135页。

[20] 孔，余，徐，伊克巴尔，王，普拉布利，弱
IEEE/ACM的注意力神经网络标记音频集
《音频、语音和语言处理学报》，第27卷，pp。
17911802，2019.

[21]A.Van Den Oord、S.Dieleman和B.Schrauwen，迁移学习
by supervised pre-training for audio-based music classification," in
国际音乐信息检索学会会议
（ISMIR），2014年，第2934页。

[22]王，弱标记复音事件检测博士
论文，卡内基梅隆大学，2018年。

[23]E.Law和L.Von Ahn，投入协议：一种新的机制
使用人类计算游戏收集数据，发表在《美国科学院院刊》上
2009年SIGCHI计算系统人为因素会议，
第11971206页。

[24]A.Mesaros、T.Heittola和T.Virtanen，TUT声学数据库
欧洲会议上的场景分类和声音事件检测-
pean信号处理会议（EUSIPCO），2016，第11281132页。

[25]J.Pons和X.Serra，MUsICNN：预训练卷积神经网络
音乐音频标签网络”arXiv预印本arXiv:1909.06654，
2019.

[26]A.Diment和T.Virtanen，“弱标记音频的迁移学习”
在IEEE音频和音频信号处理应用研讨会上
声学（WASPAA），2017，第610页。

[27]李、塞西、迪米特罗娃、麦基，《一般分类》
用于基于内容的检索的音频数据，“模式识别字母，
第22卷，第533544页，2001年。

[28]L.Vuegen，B.Broeck，P.Karsmakers，J.F.Gemmeke，B.Vanrumste，
H.Hamme，一种用于事件检测的MFCC-GMM方法
IEEE信号处理应用研讨会
音频与声学（WASPAA），2013年。

[29]A.Mesaros、T.Heittola、A.Eronen和T.Virtanen，声学事件
欧洲信号处理协会在现实生活记录中的检测-
参考文献（EUSIPCO），2010年，第12671271页。

[30]B.Uzkent、B.D.Barkana和H.Cevikalp，非言语环境
使用具有一组新特征的SVM进行声音分类，Interna-
国家创新计算、信息与控制杂志，第8卷，
第35113524页，2012年。

[31]戴，戴，屈，李，达斯，非常深卷积
IEEE国际会议上的原始波形神经网络
声学、语音和信号处理（ICASSP），2017，第421页
425.

[32]何，张，任，孙，深度残差学习
IEEE计算机视觉与模式会议上的图像识别
《认可》（CVPR），2016年，第770778页。

[33]孔，曹，伊克巴尔，徐，王，普拉布利，
用于音频标记、声音事件检测和
空间定位：DCASE 2019基线系统，arXiv预印本
arXiv:1904.034762019。

[34]A.Krizhevsky、I.Sutskever和G.E.Hinton，ImageNet分类-
深度卷积神经网络，”在Neura!
信息处理系统（NeurIPS），2012，第10971105页。

[35]K.Simonyan和A.Zisserman，非常深的卷积网络
大规模图像识别，“国际学习会议
陈述（ICLR），2015年。

[36]S.Ioffe和C.Szegedy，批处理归一化：加速深度
通过减少内部协变量转换进行网络训练，“在Intenationa!
机器学习会议（ICML），2015，第448456页。

[37]V.Nair和G.E.Hinton，修正线性单元改善了限制
玻尔兹曼机器，在国际机器学习会议上
（ICML），2010年，第807814页。

[38]N.斯里瓦斯塔瓦、G.Hinton、A.Krizhevsky、I.Sutskever和R.Salakhut-
dinov，Dropout：一种防止神经网络过度的简单方法-
fitting，《机器学习研究杂志》，第15卷，第1929页
1958，2014.

[39]林，陈，严，网络中的网络，国际
2014年离职代表会议（ICLR）。

[40]A.G.霍华德，M.ZhuB。陈，D.Kalenichenko，王，
T.Weyand、M.Andreetto和H.Adam，《移动网络：高效卷积》-
用于移动视觉应用的神经网络解决方案，arXiv预印本
arXiv:1704.048612017。

[41]桑德勒、霍华德、朱、日莫吉诺夫和陈，
MobileNetV2：IEEE中的逆残差和线性瓶颈
计算机视觉与模式识别会议（CVPR），2018，
第45104520页。

[42]J.Lee，J.Park，K.L.Kim和J.Nam，样本级深度卷积
Sound中使用原始波形进行音乐自动标记的神经网络
音乐计算会议，2017，第220226页。

[43]J.Salamon，C.Jacoby和J.P.Bello，城市数据集和分类学
声音研究，发表在ACM国际会议论文集上
多媒体，2014，第10411044页。

[44]张、西塞、多芬和帕兹，混音：超越
经验风险最小化”，在国际学习会议上
陈述（ICLR），2018年。

[45]朴，陈，张-C.Chiu，B.Zoph，E.D.Cubuk
和Q.V.Le，*SpecAugment：一种简单的数据增广方法
自动语音识别，”2019年INTERSPEECH。

[46]B.McFee，C.Raffel，D.Liang，D.P.Ellis，M.McVicar，E.Battenberg，
O.Nieto，“librosa：python中的音频和音乐信号分析”
《Python科学会议论文集》，2015年第8卷，pp。
1825.

[47]D.P.Kingma和J.Ba，“Adam：一种随机优化方法，
2015年国际学习表征会议（ICLR）。

[48]L.Ford，H.Tang，F.Grondin，J.Glass，深残差网络
对于大规模声学场景分析，INTERSPEECH，第25682572页，
2019.

[49]H.B.Sailor、D.M.Agrawal和H.A.Patil，*无监督滤波器组
基于卷积约束玻尔兹曼机的环境学习-
心理声音分类，载于《国际演讲》，2017年，第31073111页。

[50]K.J.Piczak，ESC：环境声音分类数据集，载于
ACM国际多媒体会议，2015，第10151018页。

[51]陈，刘，刘，张，严
基于多种分类器的声场景数据增强方案
建模，“DCASE2019挑战赛，技术代表，2019。

[52]郑和林，DCASE 2018的音频标记系统：
专注于标签噪声数据增强及其高效学习，”
DCASE挑战赛技术代表，2018年。

[53]T.Chen和U.Gupta，基于注意力的卷积神经网络
对于具有特征转移学习的音频事件分类，”
https://cvssp.org/projects/making_sense_of_sounds/site/assets!
挑战_摘要_图片/天翔_深圳.pdf，2018。

[54]E.Fonseca、M.Plakal、F.Font、D.P.W.Ellis、X.Favory、J.Pons和
X.Serra，用音频设备对自由声音进行通用标记-
bels：任务描述、数据集和基线，在检测研讨会上
声学场景和事件分类（DCASE），11月
2018年，第6973页。

[55]C。克罗斯，O.Bones，Y.曹，L.哈里斯，P.J.杰克逊，W.J.戴维斯，
王、考克斯和普拉布利，“环境的普遍化”-
心理声音分类：声音数据集和
IEEE声学、语音和声学国际会议上的“挑战”
信号处理（ICASSP），2019，第80828086页。

[56]C.刘，L.冯，G.刘，H.王，S.刘，自下而上-
用于音乐流派分类的cast神经网络
arXiv:1901.089282019。

[57]G.Tzanetakis和P.Cook，音频的音乐流派分类
信号，IEEE语音和音频处理汇刊，第10卷，
第293302页，2002年。

[58]曾，毛，彭，易，基于谱图的多任务
音频分类，“多媒体工具和应用，第78卷，pp。
37053722，2019.

[59]S.R.Livingstone、K.Peck和F.A.Russo，《RAVDESS:The Ryerson》
年会“情感言语与歌曲视听数据库”
加拿大脑、行为和认知科学学会
（CSBCCS），2012年，第14591462页。

孔秋强（S17）于2012年获得中国广州华南理工大学理学士学位，2015年获得硕士学位。他于2020年获得英国吉尔福德萨里大学的博士学位。在获得博士学位后，他加入字节跳动人工智能实验室担任[illegible]家。他的研究课题包括一般声音和音乐的分类、检测和分离。他以开发用于音频标[illegible]神经网络而闻名，并在2017年赢得了声学场景和事件检测和分类（DCASE）挑战[illegible]的音频标记任务。他被提名为2019年萨里大学年度研究生。他是该领域期刊和会议的常客，包括IEEE/ACM[illegible]语音和语言处理汇刊

王文武（M02-SM11）出生于中国安徽。他于1997年获得中国哈尔滨工程大学理学士学位，2000年获得[illegible]学位，2002年获得博士学位。在2007年5月加入英国萨里大学之前，他曾在伦敦国王学院[illegible]阅集团有限公司（现为Antix Labs有限公司）和创意实验室工作，目前是[illegible]处理和[illegible]学习教授，也是视觉语音和信号处理中心机器听觉实验室的共同主任。自2018[illegible]以来，他一直是中国青岛科技大学的客座教授。他目前的研究兴趣包括盲信号处理、稀疏信号处理、视听信号处理、机器学习和感知、机器听觉（听力）和统计异常检测。他在这些领域（合著）了200多篇出版物。2014年至2018年，他担任《IEEE信号处理汇刊》的副主编。他也是英国布莱顿ICASSP 2019的出版联合主席。他目前担任IEEE信号处理汇刊的高级区域编辑[illegible]IEEE/ACM音频语音和语言处理汇刊副主编。

Yin Cao（M18）于2008年获得中国南京大学电子科学与工程学士学位，2013年获得中国科学院声学研究所博士学位。随后，他在美国杨百翰大学声学组和中国科学院声学研究所工作。2018年，他[illegible]大学视觉、语音和信号处理中心。他的研究课题包括主动噪声控制、空气声学和信号[illegible]的检测、分类和分离。他以在分散主动噪声控制、加权空间梯度控制度量以及复音[illegible]音事件检测和定位方面的工作而闻名。他是2020年声学场景和事件检测和分类（DCASE）挑战赛[illegible]音标记的获胜者，并在2019年DCASE挑战赛中取得了声音事件检测和定位任务的第二好成绩。自2020年以来，他一直担任《噪声控制工程杂志》的副主编。他也是IEEE/ACM音频、语音和语言处理汇刊的常客。

Mark D.Plumbley（S88-M00-SM12-F15）分别于1984年和1991年获得英国剑桥大学电气科学学士学位和神经网络博士学位。在获得博士学位后，他成为伦敦国王学院的讲师，然后于2002年[illegible]伦敦玛丽女王大学。随后，他成为数字音乐中心的教授兼主任，并于2015年加入萨里大学，[illegible]任信号处理教授。他以分析和处理音频和音乐而闻名，使用了广泛的信号处理技术，包括矩阵分解、稀疏表示和深度学习。他是最近出版的《声音场景和事件的计算分析》一书的联合编[illegible]和最近举行的DCASE 2018声学场景和事件检测和分类研讨会的联合主席。他是IEEE信号处理学会信号处理理论和方法技术委员会的成员，也是IET和IEEE的研究员。

Turab Iqbal于2017年获得英国萨里大学电子工程学士学位，目[illegible]，他正在萨里大学视觉、语音和信号处理中心（CVSSP）攻读博[illegible]学位。在攻读博士学位期间，他使用机器学习方法在音频分类和定位领域开展了许多项目。他的研究主要集中在弱标记或噪声训练数据的学习上。