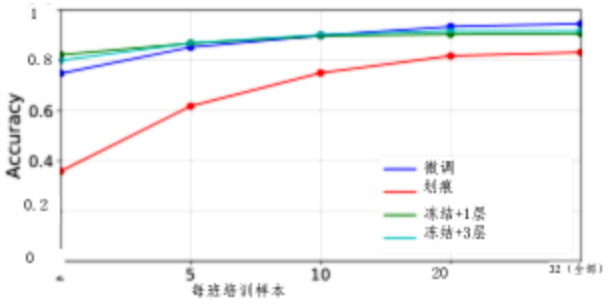

图6。ESC-50的准确性，每节课有不同数量的训练片段。

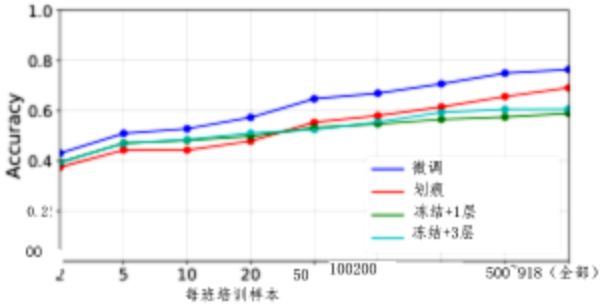

图7:DCASE 2019任务1的准确性，每节课有不同数量的训练片段。

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

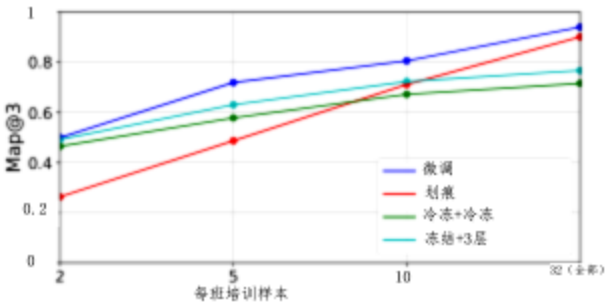

图8.DCASE 2018任务2的准确性，每节课有不同数量的训练片段。

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