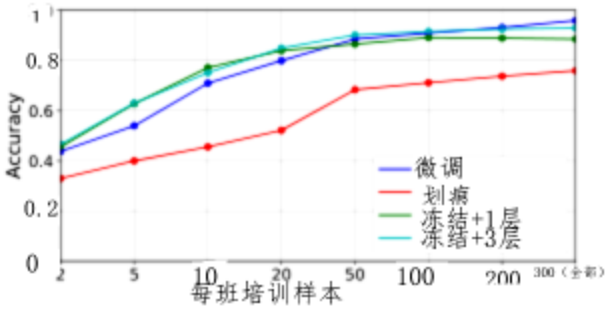

图9。MSoS的准确性，每节课有不同数量的训练片段。

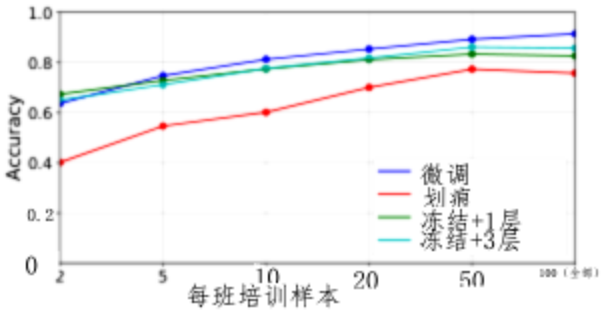

图10。GTZAN的准确性，每节课有不同数量的训练片段

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