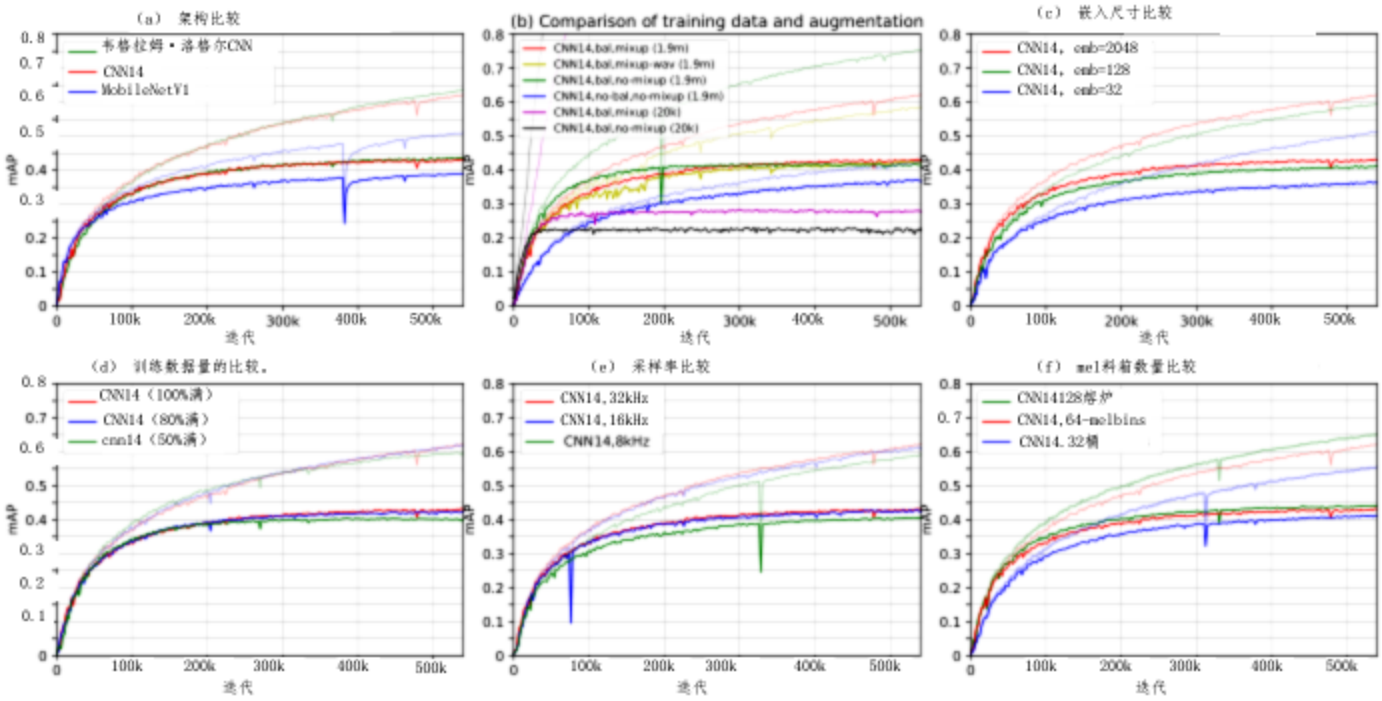

图4.音频集标记的PANNs结果。透明和实线分别是训练mAP和评估mAP。六个图显示了不同的结果：（a）架构；（b）数据平衡和数据增强；（c）嵌入尺寸；（d）训练数据量；（e）采样率；（f）梅尔垃圾桶的数量

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]。