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