arXiv:2302.07315v1 [eess.AS] 14 Feb 2023

A DAT 用于电影视听声音事件检测的 ASET

Rajat Hebbar、*Digbalay Bose*、*Krishna Somandepalli*、*Veena Vijai*、*Shrikanth Narayanan*

Si信号分析与理解实验室，南加州大学

摘要

音频事件检测是一项受到广泛研究的音频处理任务，其应用范围从自动驾驶汽车到医疗保健。Audioset 等真实环境数据集推动了该领域的研究。然而，许多研究通常需要人工标注和验证，而大规模执行这些工作成本高昂。电影描绘了各种现实生活和虚构场景，因此是挖掘广泛音频事件的丰富资源。在本研究中，我们提出了一个名为字幕对齐的电影声音（SAM-S）的音频事件数据集。我们利用公开可用的隐藏式字幕转录文本，从 430 部电影中自动挖掘出超过 11 万个音频事件。我们确定了用于对音频事件进行分类的三个维度：声音、声源、属性，并介绍了生成包含 245 种声音的最终分类体系所涉及的步骤。我们讨论了构建该分类体系时涉及的选择，同时强调了数据集中声音以人为中心的特性。对于仅使用音频的声音分类，我们建立了 34.76% 平均精度均值的基准性能，并表明融入视觉信息可以使性能进一步提升约 5%。数据和代码已公开，可通过 https://github.com/usc-sail/mica-movie-audio-events 获取，用于研究。

索引词— 音频事件检测、电影、音频识别、视听数据集

1. 引言

音频事件是自然产生的非语言声音，由人或物体发出。稳健地检测此类音频事件，可以揭示有关声学环境及个体心理状态的信息，并有助于自动生成丰富的多媒体数据转录。音频事件检测（AED）应用于广泛的领域，包括智能手机、 [1], 智能音箱 [2]和自动驾驶汽车等情境感知智能设备应用， [3, 4], 用于健康与福祉应用的声学监测 [5, 6]，以及大规模多媒体索引 [7, 8]。近年来，Audioset [7]和 VGGSound [9]等大规模“真实场景”数据集的推出，推动了音频事件检测研究的蓬勃发展。在 Audioset 上学习到的神经网络表征已用于多项音频分类任务，例如情绪识别 [10], 、性别识别 [11] 和音乐分类 [10, 12]。然而，此类数据集的整理通常需要在多个阶段进行人工干预，包括数据收集、标注和分类体系生成。此外，从 YouTube 获取的数据会因视频被下架或设为私享而逐渐流失。

作为音效，音频事件是电影音轨不可或缺的组成部分。在电影场景中，精心安排声音和背景配乐有助于构建丰富的叙事，并引发观众预期的情绪反应。电影中的大部分音效都是自然产生的（人类/动物发声、音乐等），但有些声音被称为“拟音” [13]，是在后期制作中添加的。拟音涉及



图 1。电影中显示画外音频事件的隐藏式字幕

使用“日常”物品来制作音效，模仿不同环境氛围中自然发生的音频事件；例如，折断芹菜来模仿骨头断裂的声音，或敲击垃圾桶底部来放大心跳声[1]。拟音是一种有效的工具，能够以简单、低成本的方式重现这类声音。它还可以生成上述数据源中可能不常见的音频事件（例如，车辆碰撞、轻微的脚步声、枪声）。此外，研究表明，拟音与自然产生的对应声音几乎无法区分 [14], ，因此可用于开发音频事件检测模型。

隐藏式字幕（CC）是与时间对齐的角色对话和音效转录文本。为了让媒体更便于听障人士和非母语者获取，包括电影和电视节目在内的多种广播媒体都要求提供此类字幕。根据描述与字幕媒体计划（DCMP）提供的指南， [15], 音频字幕标注员需要标注被认为与电影/电视节目剧情相关的音频事件。因此，现有字幕可用于精确获取电影数据中的音频事件。

我们的工作有三项贡献：
1. 我们采用简单且可扩展的方法，自动从电影中提取音频字幕，并从声音、声源 和属性这几个维度对音频事件进行分类。

2. 我们提出了一种与声源解耦的声音扁平分类体系。与以往的音频事件检测分类体系不同，该体系使我们能够将来自不同声源、但在声学上相似的声音归为一类。
3. 我们在 Transformer 架构中，通过早期融合音频和视频特征来利用视觉线索，从而在我们的数据集上建立音频事件检测性能基线。

本文其余部分安排如下：第 2 节讨论现有资源和方法。第 3 节概述使用字幕标签生成分类体系的过程，并与 AudioSet 分类体系进行比较。第 4 节介绍如何在我们的数据集上开发基线音频事件检测模型。

[1]博客.storyblocks.com/inspiration/拟音-sfx-everyday-household -objects

2. 相关工作

数据资源： Audioset [7], 是首批大规模音频事件检测数据集之一，包含超过 200 万个带有弱标签的片段，涵盖音频事件类别。Audioset本体是最全面的音频事件分类体系，按照音频事件的声源，以层级结构组织了 527 种不同的音频事件。Audioset 数据由人工标注员在片段级别进行标注。

为了减少标注所需的人工工作量，研究者提出了 VGGSound [9] 数据集，该数据集采用可扩展的流程，从 YouTube 中挖掘<style id='2'>具有视觉依据的音频事件。现有的机器学习模型被用于自动验证视觉特征是否存在，并在数据整理过程中排除可能的假阳性音频类别。然而，这类方法的缺点在于，仍无法保证标记的声音事件确实发生，因此只能以一定程度的标签噪声换取人工工作量的减少。此外，音频事件经常会被听到却未出现在画面中，而 VGGSound 数据集并未涵盖这种情况。

FSD50K [16] 包含从真实环境中收集的超过 5万个音频事件，这些事件在 Freesound平台 [17]上依据 200 个 Audioset 类别进行了标注。除此之外，还存在一些规模较小的音频事件检测数据集，例如 Mivia [18], 、DESED [19], 和 UrbanSound8k [20]。这些数据集通常针对特定的声音子集，例如室内、室外和罕见音频事件。

表 1。不同音频事件检测数据集的详细信息。这里，SL 表示音频事件标签是精确标注还是弱标注，PA 表示数据集是否公开；FSD 指 Freesound 平台。

数据集	领域	片段	类别	SL	PA	标注
Mivia [18]	合成数据	6K	3	✓	✓	人工标注
UrbanSound8k [20]	FSD	8.7K	10	✓	✓	人工标注
DESED [19]	FSD	12K	10	✓	✓	人工标注
FSD50K [16]	FSD	50K	200	✗	✓	人工标注
Audioset [7]	YouTube	2.1M	527	✗	✗[2]	人工标注
VGGSound [9]	YouTube	200K	309	✗	✗	自动
SAM-S	电影	110K	191	✓	✓	半自动

音频事件检测： 直到最近，卷积神经网络（CNN）模型一直广泛用于音频事件检测。VGG-16 的轻量级版本 VGGish [21] 是首个用于 Audioset 音频事件检测的基准模型。研究人员在 Audioset[10], 上比较了几种常用的卷积神经网络架构，结果表明，轻量级卷积神经网络的性能可与规模更大的同类模型相媲美。近年来，VATT [22] 和 AST [23] 等 Transformer 架构在纯音频事件检测方面展现出最先进（SoA）性能。VATT 使用自监督对比损失对通用多模态表征进行预训练，再将其微调用于音频事件检测。AST 则将最先进的视觉 Transformer [24] 应用于音频事件检测，并使用频谱图特征。

目前已有少数用于音频事件检测与定位的视听方法。研究者将中层注意力融合应用于采用卷积神经网络骨干网络的视听流 [25]。为解决不同模态间过拟合率不一致的问题，有人提出了一种名为梯度混合 [26], 的最优多模态融合机制。最近，有研究提出了多模态 Transformer 架构中的注意力瓶颈 [27], , 并通过早期融合在 Audioset 上展现出最先进的视听性能。跨模态注意力机制已用于事件声源的视听定位以及弱监督检测 [28, 29, 30]。

[2]YouTube 政策：https://www.youtube.com/static?gl=US&template=ter ms

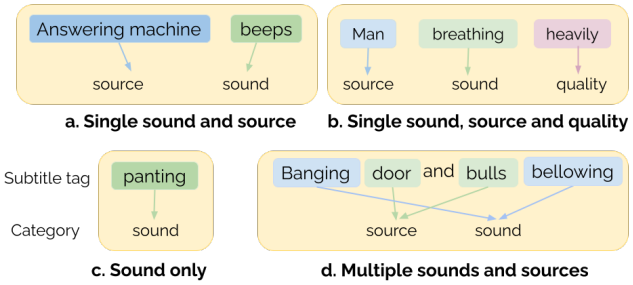


图 2。电影音频事件中声音、声源和属性类别的标注示例

在多媒体领域，分析音频事件的研究仍然有限。研究者基于 MFCC特征进行字典学习，研究了枪击声和爆炸声分类 [31]。研究者还通过聚类方法，在 8 部电影中探索了音频事件变化检测 [32]。

3. 与字幕对齐的电影声音（SAM -S)

本文介绍的 SAM-S数据集包含 430 部 2014年至2018年间票房最高的好莱坞电影。为挖掘音频事件，我们自动获取了每部电影的字幕文件[3]。以这种方式提取的隐藏式字幕包含与时间对齐的角色对白和与情节相关的事件标签。需要注意的是，隐藏式字幕并未穷尽电影中发生的所有音频事件，也就是说，这种标注方式的召回率较低。不过，就所标注的声音而言，这些字幕标签准确度很高，也就是精确率高。这种精确率—召回率权衡意味着，尽管我们会舍弃可能有用的数据，却能确保标注和清理所需的额外人工工作量最低，同时获得足够多的声音，以开发和评估音频事件检测模型。

这些标签通常用花括号括起来。我们自动从字幕文件中提取这些标签及其对应的时间戳。总计获得了略多于 11.6 万个字幕标签，其中 20,817 个为独特标签。这些字幕标签具有描述性，包含关于所发生音频事件的声音、声源和属性的信息（见图2）。虽然声音信息始终存在，但通常不会标注音频事件的声源和属性。事实上，根据 DCMP 指南，字幕标注员应标注声音的声源，但声源在屏幕上清晰可见的情况除外。在 2.1 万个独特标签中，1,500 个是一元词组，通常只表示声音；11,000 个是二元词组，包含声音和声源；其余为 n 元词组，$n \geq 3$, 可能指音频事件的属性，也可能指多个同时发生的事件。由于我们的数据中存在声源模糊的音频事件，因此我们选择采用扁平分类体系，而非像 [7], 中那样采用层级分类体系；我们将在第 3.2 节对此进行更详细的讨论。

3.1. 分类体系生成

以下步骤概述了标注、完善并凝练最终分类体系的流程：**分类：** 我们开展了一项标注任务，使用 Mechanical Turk 对给定的字幕标签进行分类，任务内容为

[3]https://github.com/ruediger/VobSub2SRT

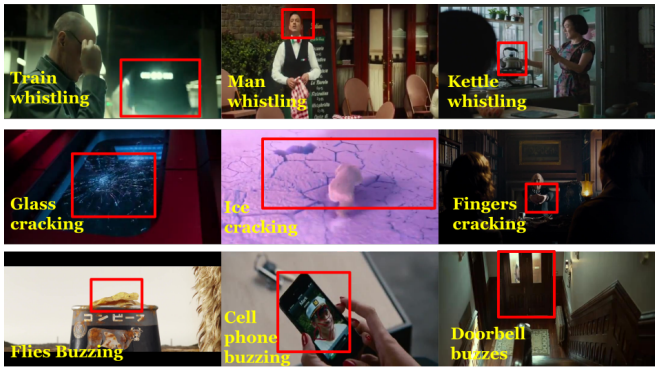


图 3. 电影中来自多个声源的声音示例：（顶部）SAM-S 中来自不同声源的“口哨声”，（中部）“噼啪声”，（底部）“嗡嗡声”

将其归入“声源”“声音”和“属性”类别。图2展示了该标注任务的几个示例。我们明确要求标注者只从标签中的词语进行选择，不得根据上下文引入或推断内容（例如，字幕标签“闪烁”只对应“闪烁”这一声音，不包含声源）。每个标签由三人标注，并对各类别采用多数表决。所有平票情况均由一位作者裁定。在这项任务中，我们选取了在数据集中至少出现5次的字幕标签。这组共2161个标签覆盖了数据集中80%（~91K）的视听事件。

词形还原： 对于每个类别，我们使用开源自然语言处理工具包 spaCy[4] 对获得的标注集合进行了词形还原。词形还原过程经过人工核验，作者也修正了其中的错误。此后，共有 254 个声源、254 种声音和 115 种属性构成了我们数据集的初始分类体系。**自动标注：** 接下来，我们创建了一个词典，将字幕标签中的原始词语映射到各类别对应的转换后版本。借助该词典，我们尝试对因出现频率较低而未纳入人工标注流程的 ~25K 个标签进行自动标注。如果同时检测到声音和声源，我们还会进行额外检查，以确保该声源组合曾出现在人工标注方案中。新的组合会被视为标注错误而忽略，相应样本也不会被使用。任何未标注声音的音频事件都会被舍弃。通过这种方式，我们又自动标注了约 10K 种声音和 5K 个声源。

标签集细化： 我们对唯一的声音和声源标签进行了最后一轮人工标注，并合并了词形还原未处理的类别，例如“笑”和“笑声”、“爆炸”和“爆炸事件”，以及“雷声”和“雷鸣”。我们使用命名实体识别器[5] 来检测人物姓名，并将其合并为一个声源类别。最终的数据集包含 95,452 个样本，涵盖来自 245 个类别的 101,311 个声音、来自 183 个类别的 21,460 个声源，以及来自 93 个类别的 7212 个属性。

3.2. 声音的声学与语义分组

我们的数据集中有若干音频事件“声音”可与多个不同的声源相关联（见图3）。例如，声音

[4]https://spacy.io/api/lemmatizer [5] https://spacy.io/usage/linguistic-features#named-entities

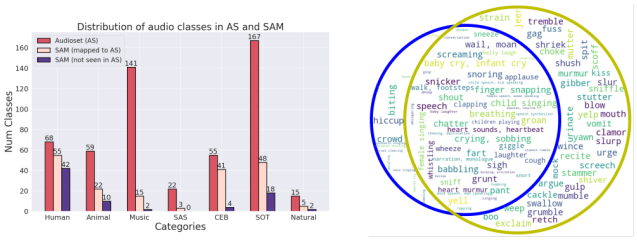


图 4。 a）Audioset 分类体系中每个类别的数量；b）Audioset（蓝色）和 SAM-S（黄色）中以人为中心的音频事件的维恩图

在我们的数据集中，嗡嗡声与不同的声源相关联——苍蝇/昆虫、手机、门铃和警报器。手机嗡嗡声和警报器嗡嗡声在声学上可以被视为相似，苍蝇和蜜蜂的嗡嗡声也是如此。然而，苍蝇/昆虫的嗡嗡声具有独特的声学特征，其频谱和时间特征使其区别于手机或警报器的嗡嗡声。从语义角度来看，这些声音通常都被称为嗡嗡声。因此，就建模而言，可以考虑以下两种方案之一：1）将特定声源声音保留为独立类别。2）合并声学和语义相似的声音

如果将这类声音视为单一类别，我们就能减少类别总数，并为每个类别获得更多具有代表性的样本，同时也会增加单个声音类别内部的声学变化。

例如，Audioset（AS）的分类体系 [7] 是层级式的，层级结构的不同分支按声源组织。在 Audioset 的五种不同层级结构中，都能看到“嗡嗡声”，分别归属于“警报器”“电话”“苍蝇”“蜜蜂”和“拟声词”。基于 Audioset 数据开发的大多数建模技术都采用方案 2，即将层级结构展开，并把每种声音视为一个单独的类别。

在我们的分类体系中，出于两个原因，我们实际选择不采用 Audioset 的方法：1. 将声源模糊的声音单独分类，会显著减少可用于训练／评估机器学习模型的样本数量；2. 根据 DCMP 指南，我们并非总能获得音频事件的声源信息。

3.3. 与 Audioset 的重叠

我们希望了解电影中声音事件的分布情况，以及它们与现有数据集的比较结果。为此，我们将数据集中的每个声音类别分为两组：a）共享声音；b）电影特有声音。共享声音是指同时存在于我们的分类体系和 Audioset 分类体系中的类别，而电影特有声音则是仅存在于我们的分类体系中的类别。

对于 SAM-S 中的每种声音，我们在 Audioset 分类体系中匹配了一个或多个对应类别 [7],，以分析标签覆盖率。我们从两个分类体系的句子转换器嵌入（MiniLM-L6-v2） [33] 之间提取了成对余弦相似度分数。随后，对每种声音在 Audioset 中排名前五的类别匹配结果进行了人工标注验证。在 245 个声音类别中，我们发现其中 170 个类别在 Audioset 中存在一个或多个直接匹配项。对于其余类别，如果确认相关，则通过人工标注将其映射到等效 AS 类别。对于源自多个可能声源的声音，我们将其分别映射到所有相关类别。最终共有

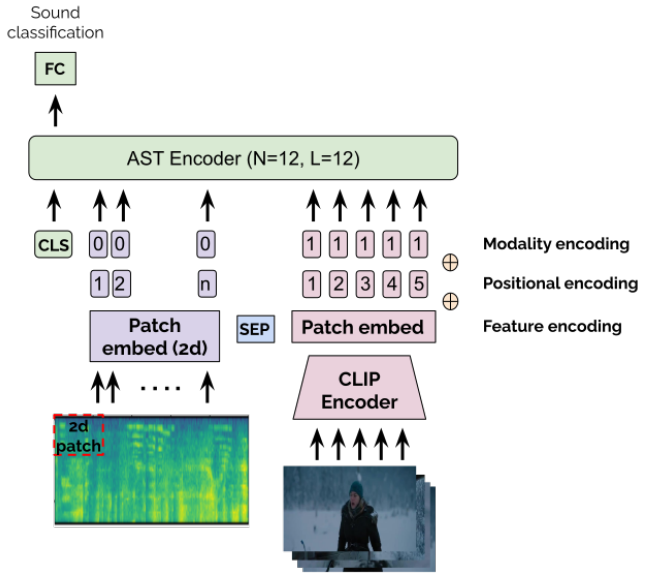


图 5。AST-MM：用于电影音频事件检测的多模态 Transformer 架构。

189 个共享声音类别。值得注意的是，尽管其中一些类别存在于 Audioset 分类体系中，但没有相应的数据样本，例如：尖叫声、刺耳声、打哈欠和喝倒彩。

最后，剩余的“电影特有声音”被人工标注并归入高层类别。“共享声音”和“电影特有声音”的分布如图 4a 所示。可以看到，人类声音和声源模糊的声音覆盖率最高。对于音乐，SAM-S 中的实例通常不会注明乐器或流派，因此我们观察不到 Audioset 中那样细致的分类。

从图 4b 可以看出，SAM-S 包含了 Audioset 中的大多数人类声音类别。不过，它还包含许多 Audioset 中没有的声音，包括打哈欠、抽鼻子、用力声和嗤笑声等单人声音，以及低语声、喧闹声和争吵声等人群声音。因此，SAM-S 也可用于扩充 Audioset，纳入更多细粒度以人为中心的声音。

4. 实验

为了从大规模模型中迁移知识，我们在音频事件的两侧各添加上下文（~5s），生成 10 秒片段。我们根据电影划分训练集、验证集和测试集（80-10-10），即使用 344 部电影进行训练、43 部进行验证，另用 43 部进行最终评估。出于开发和评估目的，我们仅保留至少占全部数据 0.1% 的声音类别，最终得到 120 个声音类别。

4.1. 基线音频模型

我们使用两种最先进模型开展纯音频基线实验。第一种是 Resnet-18 模型 [9]，使用在 VGGSound 数据集上预训练的 512 维对数频谱图。我们还微调了基于 Transformer 的音频事件检测模型 AST [23]，该模型在 Audioset 上的音频事件检测任务中展现了最先进的性能。AST 使用 128 维对数梅尔

表 2。SAM-C 上的单模态和多模态结果

Model	Modality	mAP	mAUC d	-prime
VGGSound	A	14.1	87	1.59
AST	A	34.76	95.02	2.33
AST-MM (S)	AV	35.67	95.05	2.33
AST-MM (B)	AV	35.82	95.11	2.34
AST-MM (L)	AV	**36.3**	**95.25**	**2.36**

声谱图作为特征。数据增强方面，我们以 0.5 的概率使用 mixup [34]，并从参数为 α=β=10 的贝塔分布中采样 mixup 的 lambda 值。我们还使用 SpecAug- ment [35]，其时频掩码大小为 192x48。

在实验中，我们采用批量大小为 20、初始学习率为 1e-5，并在第 5 和第 25 个 epoch 使用多步学习率调度器，衰减率为 0.85，与 [23] 类似。每个模型训练 30 个 epoch。由于我们的类别是多标签的，并且使用了 mix-up，因此采用二元交叉熵损失。评估指标包括平均精度均值（mAP）、曲线下面积（mAUC）和 d'。

4.2. AST-MM

作为视觉特征，我们使用 CLIP编码器（ViT-B/32）的输出 [36]。CLIP 通过语言-图像对以对比方式进行训练，并已广泛应用于多种图像识别任务。我们还选择 CLIP，是因为它能够很好地泛化到未见过的物体和场景，而这在电影中很常见。我们以 1 fps 提取 512 维 CLIP特征，并对得到的特征进行填充或裁剪，使其序列长度为 12。

与先前工作[37]一样，我们使用位置嵌入和特定于模态的片段嵌入，在 Transformer 架构中对多种模态进行编码。我们将 CLIP特征通过一个线性层，使其维度与音频补丁嵌入（768 维）相匹配。我们尝试了三种不同的位置嵌入：1. 固定的正弦位置嵌入（S） [38], 2。使用预训练的 BERT位置嵌入初始化的可学习嵌入（B） [39], 3。可学习的随机初始化嵌入。

我们还使用分隔符标记来区分音频和视觉序列，因为实验发现这样做有助于提升效果。通过将每个序列的补丁嵌入、位置嵌入和模态嵌入相加，得到输入编码器的最终表示。

4.3. 结果

从表 2 可以看出，在 Audioset 上训练的纯音频模型明显优于在 VGGSound 上训练的模型。除了数据集规模和模型架构之外，这可能是因为视听对应关系的约束限制了 VGGSound 中涵盖的声音类别范围，从而影响了它向其他领域的迁移能力。多模态 AST-MM 模型相较于纯音频模型提升了 5%。

5. 结论

本文发布了一个专为电影音频事件检测整理的数据集。我们介绍了一种生成音频事件平面分类体系的可扩展方法，并将其与音频事件检测中广泛使用的现有分类体系进行了比较。我们设计了一套标注方案，仅依据字幕标签对音频事件的声音和声源进行分类。我们采用最先进的机器学习模型，在 SAM-S 语料库上建立音频事件检测性能基线。

我们以早期融合的方式利用 CLIP编码器特征中的视觉信息，进一步提升多模态音频事件检测性能。

6. 参考文献

[1] “探索 soundanalysis 中内置的声音分类功能”，.[2] “识别音频流中的声音”，· [3] Mahesh Kumar Nandwana 和 Taufiq Hasan，“迈向能够聆听的智能汽车，道路上的异常声学事件检测。”，载于 *INTER- SPEECH*，2016年，第2968–2971页。[4] Annamaria Mesaros、Toni Heittola、Aleksandr Diment、Benjamin Elizalde、Ankit Shah、Emmanuel Vincent、Bhiksha Raj 和 Tuomas Virtanen，“DCASE 2017挑战赛设置：任务、数据集和基线系统”，载于 *DCASE 2017-声学场景与事件检测研讨会*，2017年。[5] Arindam Jati、Amrutha Nadarajan、Raghuveer Peri、Karel Mundnich、Tiantian Feng、Benjamin Girault 和 Shrikanth Narayanan，“工作场所声学场景的时间动态：自我和环境分析与预测”，*IEEE/ACM音频、语音与语言处理汇刊*，第29卷，第756–769页，2021年。[6] Stefan Goetze、Jens Schroder、Stephan Gerlach、Danilo Hollosi、Jens-E Appell 和 Frank Wallhoff，“社会照护中的声学监测与定位”，计算科学与工程杂志，第6卷，第1期，第40–50页，2012年。[7] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter，“Audio Set：一个用于音频事件的本体与人工标注数据集”，载于 *2017年IEEE国际声学、语音与信号处理会议（ICASSP）*。IEEE，2017年，第776–780页。[8] Benjamin Ma、Timothy Greer、Dillon Knox 和 Shrikanth Narayanan，“透过计算视角探究音乐如何塑造电影流派”，*PloSone*，第16卷，第4期，第e0249957页，2021年。[9] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman，“VGGSound：大规模视听数据集”，载于 *ICASSP 2020-2020 IEEE国际声学、语音与信号处理会议（ICASSP）*。IEEE，2020年，第721–725页。[10] Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley，“PANNs：用于音频模式识别的大规模预训练音频神经网络”，*IEEE/ACM音频、语音与语言处理汇刊*，第28卷，第2880–2894页，2020年。[11] Rajat Hebbar、Krishna Somandepalli 和 Shrikanth S Narayanan，“利用跨域数据提升电影音频单的性别识别能力。”，载于 *INTERSPEECH*，2018年，第282–286页。[12] Jaime Herm´ırez 和 M. Julia Flores，“音乐流派的机器学习：结合 Audioset 的多面向探述与实验”，智能信息系统杂志，第1–31页，2019年。[13] Claire Nozaic，电影音频后期制作简介，博士论文，斯泰伦博斯：斯泰伦博斯大学，2006年。[14] Antalia De G¨otzen、Erik Sikstr¨om、Francesco Grani 和 Stefania Ser- afin，“真实、拟音还是合成数据？日常行走声音评估”，*SMC论文集*，2013年。[15]“电视和电影中的音效字幕——3Play Media”，https://www.3playmedia.com/blog/captioning-sound-effects-in-tv-and-movies/.[16] Eduardo Fonseca、Xavier Favory、Jordi Pons、Frederic Font 和 Xavier Serra，“FSD50K：一个人工标注声音事件的开放数据集”，*IEEE/ACM音频、语音与语言处理汇刊*，第30卷，第829–852页，2022年。[17] Eduardo Fonseca、Jordi Pons Puig、Xavier Favory、Frederic Font Cor- bera、Dmitry Bogdanov、Andres Ferraro、Sergio Oramas、Alastair Porter 和 Xavier Serra，“Freesound数据集：开放音频数据集创建平台”，载于 第*18*届*ISMIR*会议论文集；第*486–493*页。，2017年。[18] Pasquale Foggia、Nicolai Petkov、Alessia Saggese、Nicola Strisciuglio 和 Mario Vento，“ 高噪声环境中可靠的音频事件检测 ”，*PatternRecognitionLetters*，第65卷，第22–28页，2015年。

[19] Nicolas Turpault、Romain Serizel、Justin Salamon 和 Ankit Parag Shah，《利用弱标注数据和声景合成进行家庭环境中的声音事件检测》，载于《声学场景与事件检测和分类 2019 研讨会论文集》（DCASE2019），2019年，第253–257页。[20] Justin Salamon、Christopher Jacoby 和 Juan Pablo Bello，《城市声音研究的数据集与分类体系》，载于第22届 ACM 多媒体国际会议论文集，2014年，第1041–1044页。[21] Shawn Hershey、Sourish Chaudhuri、Daniel PW Ellis、Jort F Gemmeke、Aren Jansen、R Channing Moore、Manoj Plakal、Devin Platt、Rif A Saurous、Bryan Seybold 等，《用于大规模音频分类的卷积神经网络架构》，载于2017年 IEEE国际声学、语音与信号处理会议（ICASSP），IEEE，2017年，第131–135页。[22] Hassan Akbari、Liangzhe Yuan、Rui Qian、Wei-Hong Chuang、Shih-Fu Chang、Yin Cui 和 Boqing Gong，《VATT：利用原始视频、音频和文本进行多模态自监督学习的Transformer》，《神经信息处理系统进展》，第34卷，2021年。[23] Yuan Gong、Yu-An Chung 和 James Glass，《AST：音频频谱图Transformer》，arXiv预印本 arXiv:2104.01778，2021年。[24] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Herv´e J´egou，《通过注意力进行数据高效图像Transformer训练与蒸馏》，载于国际机器学习会议。PMLR，2021年，第10347–10357页。[25] Haytham M Fayek 和 Anurag Kumar，《利用弱标注数据进行大规模视听声音学习》，arXiv预印本 arXiv:2006.01595，2020年。[26] Weiyao Wang、Du Tran 和 Matt Feiszli，《是什么让多模态分类网络的训练变得困难？》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第12695–12705页。[27] Arsha Nagrani、Shan Yang、Anurag Arnab、Aren Jansen、Cordelia Schmid 和 Chen Sun，《用于多模态融合的注意力瓶颈》，《神经信息处理系统进展》，第34卷，2021年。[28] Mathilde Brousmiche、St´ephane Dupont 和 Jean Rout，《用于视听事件检测的模态内与模态间交互》，载于第1届以人为中心的多媒体分析国际研讨会论文集，2020年，第5–11页。[29] Yu Wu、Linchao Zhu、Yan Yan 和 Yi Yang，《用于视听事件定位的双重注意力匹配》，载于 IEEE/CVF国际计算机视觉会议论文集，2019年，第6292–6300页。[30] Hanyu Xuan、Zhenyu Zhang、Shuo Chen、Jian Yang 和 Yan Yan，《用于时间不一致视听事件定位的跨模态注意力网络》，载于AAAI人工智能会议论文集，2020年，第34卷，第279–286页。[31] C´edric Penet、Claire-H´el`ene Demarty、Guillaume Gravier 和 Patrick Gros，《利用多个音频词和上下文贝叶斯网络检测电影中的音频事件》，2013年第11届基于内容的多媒体索引国际研讨会（CBMI），第17–22页，2013年。[32] Ji-chen Yang、Lei-an Liu、Qing-wei Qin 和 Min Zhang，《电影中的音频事件变化检测与聚类》，《多媒体学报》，第8卷，第2期，第113页，2013年。[33] Nils Reimers 和 Iryna Gurevych，《Sentence-BERT：使用孪生BERT网络生成句子嵌入》，载于2019年自然语言处理实证方法会议论文集。第2019年11月，计算语言学协会。[34] Yuji Tokozume、Yoshitaka Ushiku 和 Tatsuya Harada，《利用类间样本进行深度声音识别学习》，arXiv预印本 arXiv:1711.10282，2017年。[35] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Bar-ret Zoph、Ekin D Cubuk 和 Quoc V Le，《SpecAugment：一种用于自动语音识别的简单数据增强方法》，arXiv预印本 arXiv:1904.08779，2019年。[36] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever，《通过自然语言监督学习可迁移视觉模型》，载于ICML，2021年。

[37] Douwe Kiela、Suvrat Bhooshan、Hamed Firooz、Ethan Perez 和 Da-vide Testuggine，《用于图像与文本分类的监督式多模态双向变换器》，*arXiv*预印本 *arXiv:1909.02950*，2019年。[38] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan NGomez、Lukasz Kaiser 和Illia Polosukhin，《注意力机制就是你所需要的一切》，神经信息处理系统进展，第30卷，2017年。[39] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，*arXiv*预印本 *arXiv:1810.04805*，2018年。