arXiv:2303.03689v1 [eess.AS] 7 Mar 2023

AST-SED：一种基于音频频谱图变换器的有效声音事件检测方法

Kang Li[1], *Yan Song*[1], *Li-Rong Dai*[1],*Ian McLoughlin*[1,2],*Xin Fang*[3],*Lin Liu*[3]

[1]国家语音及语言信息处理工程研究中心，中国科学技术大学，中国合肥。[2]ICT 集群，新加坡理工大学，新加坡。[3]科大讯飞研究院，科大讯飞股份有限公司，中国合肥。

摘要

本文提出了一种有效的声音事件检测（SED）方法，该方法基于在大规模 AudioSet 音频标注（AT）任务上预训练的音频频谱图变换器（AST）模型，称为 AST-SED。预训练 AST 模型近期在 DCASE2022 挑战赛任务 4 中展现出良好效果，有助于缓解真实标注数据不足的问题。然而，主要由于 AT 任务与 SED 任务之间存在差异，直接利用预训练 AST 模型的输出并非最优。因此，所提出的 AST-SED 采用编码器-解码器架构，无需重新设计或重新训练 AST 模型，即可实现有效且高效的微调。具体而言，频率维度变换器编码器（FTE）沿频率轴采用带自注意力的变换器，以解决单个音频片段中多个音频事件重叠的问题。局部门控循环单元解码器（LGD）由最近邻插值（NNI）和双向门控循环单元（Bi-GRU）组成，用于补偿预训练 AST 模型输出中的时间分辨率损失。在 DCASE2022 任务 4 开发集上的实验结果表明，采用 FTE-LGD 架构的 AST-SED 性能优越。具体而言，其基于事件的 F1 分数（EB-F1）达到 59.60%，多声道声音检测评分场景 1（PSDS1）得分为 0.5140，显著优于 CRNN 及其他基于预训练 AST 的系统。

索引词— 声音事件检测，Transformer，门控循环单元，编码器-解码器，微调

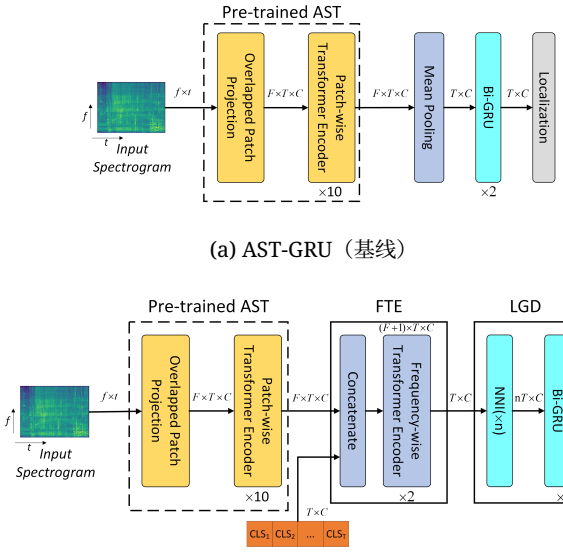
1. 引言

声音事件检测（SED）旨在确定给定音频片段中多个重叠事件的类别和时间戳。它广泛应用于现实世界的系统，包括智能家居设备 [1] 和自动监控 [2]。最近，DCASE 挑战赛设置了一系列任务[1] ，用于评估声音事件检测的研究进展——现有系统通常采用半监督学习方法，例如用于弱标注声音事件检测任务的 Mean Teacher [3, 4, 5], 。其中大多数系统利用卷积循环神经网络（CRNN）架构 [6, 7] 及其变体，例如 SK-CRNN [8] 和 FDY-CRNN [9] ，来执行声音事件检测的帧级特征提取和上下文建模。2020年，Miyazaki 等人 [10] 引入了 Conformer——一种卷积增强型 Transformer 架构，用于同时建模局部和全局上下文信息。借助强标注数据和合成数据，声音事件检测性能可以得到进一步提升。

在 DCASE2022 中，有多项工作提出利用外部的大规模弱标注 AudioSet [11] 数据。例如，Ebbers 和 Haeb-Umbach 的前向-后向卷积循环神经网络（FB-CRNN）和双向 CRNN（Bi-CRNN） [12], 首先在 AudioSet 上进行预训练，然后通过自训练方式针对声音事件检测进行微调。Xiao [13] 使用循环神经网络对预训练音频神经网络（PANN）和音频频谱图变换器（AST）的输出进行上下文建模; 这两种模型 [14, 15], 最初都是为音频标注任务设计的。受 [13],启发，我们使用预训练 AST 模型扩展 AST-GRU，以用于声音事件检测，如图1(a)所示。与卷积循环神经网络类似，首先沿频率轴对预训练 AST 的输出进行聚合（均值池化），形成帧级序列，然后将其输入双向门控循环单元进行上下文建模。

尽管 AST-GRU 已取得令人瞩目的性能，但如上所述，由于音频标记任务与声音事件检测任务之间存在差异，直接使用预训练 AST 的输出仍不是最优方案。首先，单个音频片段中会出现多个具有不同频带的音频事件，因此简单的聚合方案（例如 平均池化）可能会造成频率信息丢失。其次，AST 模型是在 AudioSet 上针对音频分类任务进行预训练的，侧重于音频片段级表示，因此可能导致时间分辨率损失。本文提出一种有效的 AST-SED 方法，该方法采用基于在大规模 AudioSet 上预训练的 AST 模型的编码器-解码器架构，如图 1(b) 所示。具体而言，在频率维度变换器

[1]https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments



(a) AST-GRU（基线）

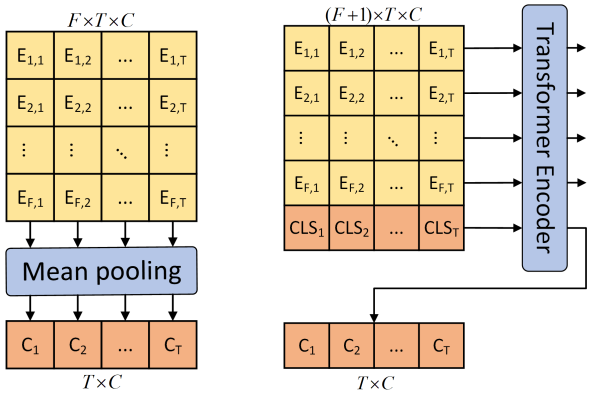
(b) AST-SED（我们提出的模型）

图 1。(a) AST-GRU：沿频率轴进行平均池化，从 AST 的输出中获得帧级序列，并将其输入双向门控循环单元进行上下文建模。(b) AST-SED：频率维度变换器编码器（FTE）生成帧级序列，随后采用最近邻插值（NNI），并通过双向门控循环单元构成局部GRU解码器（LGD），以生成高时间分辨率特征。

编码器（FTE）中，将帧级类别标记与 AST 输出拼接，然后输入一系列 Transformer 模块。沿频率轴执行自注意力，以获得适用于多个重叠事件的帧级表示。局部 GRU 解码器（LGD）首先通过最近邻插值（NNI）扩展帧序列，使其与输入频谱图的时间分辨率相匹配，随后使用双向门控循环单元 [16] 对声音事件检测任务进行上下文建模。与以往系统一样， [4], 采用 Mean Teacher 从弱标注数据和无标注数据中学习。我们在 DCASE2022 挑战赛任务 4 开发集上进行了大量实验，以评估所提出的采用 FTE-LGD 架构的 AST-SED。具体而言，该方法取得了 59.60% 的 EB-F1 和 0.5140 的场景 1 复音声音检测分数（PSDS1），优于 DCASE2022 获胜系统的 57.30% 和 0.5050。

2. 方法

本节简要介绍采用预训练AST的基线系统AST-GRU，并分析其不足之处。随后详细介绍所提出的编码器-解码器（即FTE-LGD）架构AST-SED系统，该系统无需重新设计AST模型即可实现高效微调。



(a) 平均池化

(b) 采用频率维度自注意力的 FTE 注意力

图 2。AST-GRU 与 FTE 中平均池化的比较。(a) 沿频率轴进行平均池化；(b) FTE 沿频率轴进行自注意力。

2.1. 使用预训练 AST 的基线系统：AST-GRU

在 [15, 17], 中，提出了一种用于音频分类任务的AST架构。AST-GRU系统基于预训练AST的输出构建，如图1(a)所示。输入频谱图被划分为大小为 16×16 、步长为 10×10的补丁。这些补丁经过线性投影，转换为768维的补丁标记。根据 [18, 19], ，补丁标记与位置嵌入构成输入序列**PI** $\in \mathbb{R}^{F\times T\times C}$。补丁级变换器编码器（PTE）通过多头自注意力（MHSA）进行补丁级上下文建模，如图1(a)所示。

众所周知，训练基于 Transformer 的模型通常需要大量计算资源，例如，自注意力的计算复杂度会随补丁序列的长度增加$O(n^2)$，并且需要充足的训练数据 [19]。这可能会限制基于 Transformer 的模型直接用于声音事件检测任务，即使使用大规模弱标注 AudioSet，也会受到序列长度n的影响。在 AST-GRU 中，首先沿频率轴进行平均池化，以形成长度大幅缩短n的帧序列，然后使用双向门控循环单元执行声音事件检测任务。使用带 Sigmoid 激活函数的线性分类器生成帧级声音事件检测预测，并将 softmax [20] 用于音频标签任务。此外，PTE 部分使用预训练的 *AST-base*进行初始化，该模型包含 10 个 Transformer 模块。随后，使用 Mean Teacher 对 AST-GRU 进行微调，以处理弱标注数据和无标注数据 [3]，其损失函数定义如下：

$$L_{total}(t) = 0.5 \cdot L_c + L_f + \alpha(t) \cdot (L_{MT,c} + L_{MT,f}) \quad (1)$$

其中， L_c、 L_f 分别表示音频片段级和帧级分类损失，而$L_{MT,c}$、 $L_{MT,f}$ 分别表示音频片段级和帧级师生一致性损失。权重 $\alpha(t)$

在每次迭代中使用渐增方案进行调节 t。尽管 AST-GRU 在声音事件检测方面优于现有的卷积循环神经网络架构，但由于前文所述音频标记任务与声音事件检测任务之间的差异，它仍未达到最优。下一小节将详细介绍所提出的采用 FTE-LGD 编码器-解码器架构的 AST-SED 系统，以解决 AST-GRU 中频率和时间信息的丢失问题。

2.2. 采用所提出的 FTE-LGD 架构的 AST-SED

2.2.1. FTE：频率维度 *Transformer* 编码器

我们首先评估每个事件的频率维度激活分布，以说明 FTE 的设计动机。具体而言，将 AST-GRU 中的平均池化替换为单个频带，以评估每个事件类别的检测性能。计算使用不同频带时各类别的 EB-F1，并将其归一化至 $[0,1]$，以反映频率维度激活分布。图 3 展示了四类事件的频率维度分布直方图：(1) 主要在高频带上激活的闹钟铃响事件；(2) 主要在低频带上激活的语音事件；(3) 狗、餐具和猫事件等，主要在中频段激活；以及 (4)电动剃须刀、搅拌机、吸尘器和油炸事件等，在所有频带上均会激活。我们观察到这些事件在频率维度上呈现出不同的分布，这表明平均池化等简单的聚合方案可能不可避免地丢失对声音事件检测具有潜在判别性的频率信息。

为解决这一问题，我们提出了一个 FTE模块，在 Transformer 中引入频率维度多头自注意力（fMHSA）。具体而言，给定 PTE 的输出 $\mathbf{PO} \in \mathbb{R}^{F\times T\times C}$，将一个可学习的 **CLS**token 序列 $\mathbf{CLS} \in \mathbb{R}^{T\times C}$ 沿频率轴与 **PO** 拼接，形成 FTE 的输入 $\mathbf{FI} \in \mathbb{R}^{(F+1)\times T\times C}$。接下来，[18, 19], 一个 FTE模块主要由带有 fMHSA 的多层感知机（MLP）组成，后接层归一化（LN），如下所示：

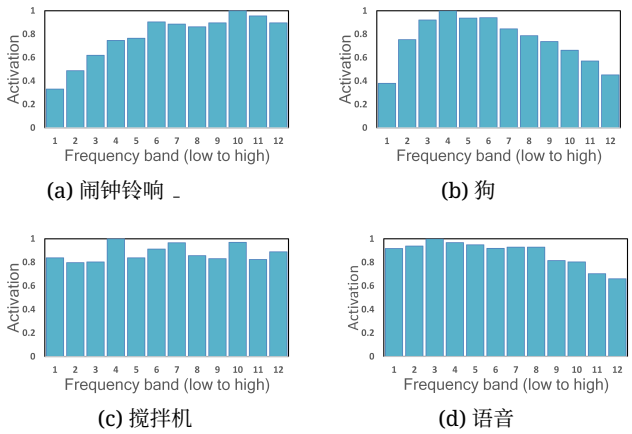
$$\hat{\mathbf{FI}} = LN(fMHSA(\mathbf{FI}) + \mathbf{FI}) \tag{2}$$

$$\mathbf{FO} = LN(MLP(\hat{\mathbf{FI}}) + \hat{\mathbf{FI}}) \tag{3}$$

需要注意的是，在 fMHSA 中，每个 token $E_{t,f}$仅与其他 token $E_{t,\cdot}$交互，如图 2(b) 所示。FTE 输出 **CLS** token $\mathbf{FO} \in \mathbb{R}^{(F+1)\times T\times C}$构成最终的帧序列 $\mathbf{C} \in \mathbb{R}^{T\times C}$。FTE 由 2 个具有 4 头 fMHSA 的 Transformer 模块组成，中间嵌入维度设为 768。

2.2.2. LGD：局部*GRU*解码器

由于 AST 输出的时间分辨率较低，直接将其输入双向门控循环单元进行上下文建模，可能只能获得有限的声音事件检测性能。因此，LGD 包含一个 NNI模块和双向门控循环单元，构成一个局部解码器，并将其应用于



(a) 闹钟铃响 (b) 狗

(c) 搅拌机 (d) 语音

图3。沿频率维度的归一化激活分布。

FTE 输出序列 $\mathbf{C} \in \mathbb{R}^{T\times C}$，如图 1(b) 所示。若 NNI 的上采样比（UR）记为 n，则 LGD 的一种简单实现如下：

$$\hat{\mathbf{C}} = NNI(\mathbf{C}, n) \tag{4}$$

$$\mathbf{O} = BiGRU(\hat{\mathbf{C}}) \tag{5}$$

该输出 $\mathbf{O} \in \mathbb{R}^{nT\times C}$ 可能适用于具有高时间分辨率的细粒度声音事件检测任务。由于卷积神经网络部分旨在生成高时间分辨率的输出，因此无需解码；而在卷积循环神经网络中，门控循环单元的解码能力 [16] 在很大程度上被忽视了。然而，在基于 AST 的声音事件检测中，由于需要大量计算资源（包括标注数据和计算成本），使用高时间分辨率的 Transformer 进行预训练或微调并不高效 [19]。这也是 Conformer [10]选择用双向门控循环单元进行声音事件检测，而非使用 Transformer 的原因。

3. 实验与结果

3.1. 数据集与实验设置

在评估中，我们采用 DCASE2022 任务 4 开发集（DESED）[21]。训练数据集包含 1578 个弱标注音频片段、14412 个无标注音频片段、3470 个真实强标注音频片段和 10000 个合成强标注音频片段。验证数据集由 1168 个音频片段组成。我们使用 PSDS1 [22] 和 EB-F1 分数评估细粒度声音事件检测性能。

首先将 16kHz 音频输入波形转换为 128 维对数梅尔频谱图特征，窗长为 25ms，帧移为 10ms，然后将其归一化，使均值为 0、标准差为 0.5。使用 MixUp [23], 时间掩蔽、时间平移和 FilterAugment [24]等数据增强方法。后处理中使用固定时长的中值滤波器，并为所有事件类型采用 0.5 的统一阈值，以获得用于计算 EB-F1 的硬预测。数据增强和中值滤波器参数与 [9] 相同。

表1。DCASE 2022 DESED 上的模型性能比较。† 表示使用来自 [9] 的代码库进行实现所得的结果。

模型	EB-F1，%	PSDS1
卷积循环神经网络†	50.50	0.4006
FDY-CRNN† [9]	51.56	0.4256
SK-CRNN† [8]	52.77	0.4004
集成 PANNs-RNN [13]	N/A	0.4450
集成 AST-RNN [13]	N/A	0.4590
双向卷积循环神经网络（获胜者） [12]	57.30	0.5050
AST-GRU† （本文方法）	**55.20**	**0.4537**
AST-SED† （本文方法）	**59.60**	**0.5140**

表2。所提出方法在 DCASE 2022年 DESED 上的性能。UR 表示上采样率。

模型	EB-F1，%	PSDS1
AST-GRU	55.20	0.4537
AST-GRU+FTE	57.86	0.4742
AST-GRU+LGD	57.50	0.4933
AST-GRU+FTE+LGD（AST-SED）	**59.60**	**0.5140**
AST-SED（UR=1）	57.86	0.4742
AST-SED（UR=2）	59.12	0.5016
AST-SED（UR=5）	**60.00**	0.5110
AST-SED（UR=10）	59.60	**0.5140**

AST-GRU 和 AST-SED 均使用 AdamW优化器 [25], 训练 10 个 epoch，每个批次中真实强标注、真实合成、弱标注和无标注数据的比例为 1:1:2:2。预训练AST 和其余部分的学习率（lr）分别设为 5e-6 和 1e-4。训练期间，前 5 个 epoch 的 lr 保持不变，之后按指数方式衰减。使用 Mean Teacher SSL，基于 DCASE2022 DESED 对 AST-GRU 和 AST-SED 进行微调。

3.2. AST-GRU 与 AST-SED 性能比较

如表1所示，AST-GRU 在 DCASE 2022 DESED 上的 EB-F1 达到 55.20%，PSDS1 达到 0.4537，显著优于卷积循环神经网络基线的相应得分 50.5% 和 0.4006。这表明，在相同实验设置下，经过 AudioSet 预训练的 AST 是一种强大的特征提取器。采用 FTE-LGD 编码器-解码器架构的 AST-SED 能取得更好的结果。具体而言，其 EB-F1 最高可达 59.60%，PSDS1 最高可达 0.5140，超过了近期报告的最先进的结果。值得注意的是，此前的获胜系统 [12] 采用自训练方式进行训练。

3.3. AST-SED与FTE-LGD消融研究

我们进行了一系列消融研究实验，以评估不同架构配置和参数选择的有效性。表2列出了采用 FTE 或 LGD 的配置，这些配置的表现可能优于基线 AST-GRU。采用 FTE-LGD 编码器-解码器架构的 AST-GRU/SED 性能最佳，相较于 AST-GRU，EB-F1 绝对提升了 4.40%，PSDS1 提升了 0.06。我们还在表2中考察了不同上采样比率下 LGD 的性能，结果表明比率为 5 和 10 时表现最佳。

表3。采用LGD时各个事件的相对提升（RI）；S：语音，D：狗，R：流水，E：电动剃须刀和牙刷，A：闹钟铃响，C：猫，F：煎炸，B：搅拌机，V：吸尘器，Di：餐具。另对短、长事件集进行了比较。

事件	Mean 时长，s	EB-F1，% （不含）	EB-F1，% (w)	RI，%
S	1.50	60.27	66.07	**9.62**
D	1.37	42.30	46.37	**9.61**
R	5.15	56.73	56.80	0.12
E	7.90	71.27	74.53	4.58
A	1.96	54.77	55.23	0.85
C	1.39	52.37	54.00	3.12
F	8.23	67.83	71.73	5.75
B	5.15	61.70	64.60	4.70
V	8.45	75.30	69.80	-7.30
Di	0.63	36.07	36.80	2.03
短	0-2	49.15	51.69	**5.17**
Long	2-10	66.57	67.49	1.39
All	-	57.86	59.59	3.00

表3区分了不同音频事件类型的性能。对于持续时间较短的事件，例如语音 和狗，AST-SED分别取得了9.62%和9.61%的显著提升。对于持续时间较长的事件，例如流水 和吸尘器，AST-SED带来的提升则较为有限。这可能是由于LGD中的NNI操作相对简单。未来，我们计划探索更强大的解码器，以更好地提升较长持续时间事件的检测效果。

4. 结论

本文提出了一种基于预训练 AST 模型的有效 AST-SED 方法。具体而言，本文设计了一种无需修改 AST 模型的编码器-解码器架构，以实现高效训练。编码器 FTE 由采用频率维度自注意力的 Transformer 模块组成，用于有效进行帧级特征提取。同时，解码器 LGD 由带有 NNI 操作的双向 GRU 模块组成，用于实现 SED 的细粒度能力。DCASE2022 任务 4 上的实验结果表明，所提出的采用 FTE-LGD 架构的 AST-SED 能够优于最先进方法。

5. 参考文献

[1] A. Southern、F. Stevens 和 D. Murphy，“为智慧城市发声：环境声音设计中的可听化与声景监测”，美国声学学会期刊，第 141 卷，第 5 期，第 3880–3880 页，2017年。[2] R. Radhakrishnan、A. Divakaran 和 A. Smaragdis，“面向监控应用的音频分析”，载于 *IEEE WASPAA 2005*。，2005年，第 158–161 页。[3] A. Tarvainen 和 H. Valpola，“平均教师是更好的榜样：权重平均一致性目标能够改善半监督深度学习结果”，神经信息处理系统进展，第 30 卷，2017年。[4]X. Zheng、Y. Song、L.-R. Dai、I. McLoughlin 和 L. Liu，“一种基于互助平均教师的有效声音事件检测领域自适应方法”，载于*Interspeech*，2021年，第 556–560 页。[5] H. Endo 和 H. Nishizaki，“用于多音声音事件检测的同伴协作学习”，载于 *IEEEICASSP*，2022年，第 826–830 页。[6] E. Cakır、G. Parascandolo、T. Heittola、H. Huttunen 和 T. Virtanen，“用于多音声音事件检测的卷积循环神经网络”，*IEEE* 音频、语音与语言处理汇刊，第 25 卷，第 6 期，第 1291–1303 页，2017年。[7] S. Adavanne、P. Pertil¨a 和 T. Virtanen，“使用空间特征和卷积循环神经网络进行声音事件检测”，载于 *IEEE ICASSP*，2017年，第 771–775 页。[8] X. Zheng、Y. Song、I. McLoughlin、L. Liu 和 L.-R. Dai，“一种改进的基于平均教师的大规模弱标注半监督声音事件检测方法”，载于 *IEEEICASSP*，2021年，第 356–360 页。[9] H. Nam、S.-H. Kim、B.-Y. Ko 和 Y.-H. Park，“频率动态卷积：用于声音事件检测的频率自适应模式识别”，*arXiv* 预印本 *arXiv:2203.15296*，2022年。[10] K. Miyazaki、T. Komatsu、T. Hayashi、S. Watanabe、T. Toda 和 K. Takeda，“用于半监督声音事件检测的卷积增强 Transformer”，载于检测与分类声学场景和事件研讨会论文集（*DCASE*），2020年，第 100–104 页。[11] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，“Audio Set：面向音频事件的本体与人工标注数据集”，载于*IEEE ICASSP*，2017年，第 776–780 页。[12] J. Ebbers 和 R. Haeb-Umbach，“面向家庭环境声音事件检测的预训练与自训练”

[13] S. Xiao，《用于 DCASE 2022 挑战赛任务 4 的声音事件检测预训练模型》，技术报告，DCASE，2022年6月。[14] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang 和 M. D. Plumbley，《用于音频模式识别的大规模预训练音频神经网络》，*IEEE* 音频、语音与语言处理汇刊，第 28 卷，第 2880–2894 页，2020年。[15] Y. Gong、Y.-A. Chung 和 J. Glass，《AST：音频频谱图变换器》，载于 *Interspeech 2021* 论文集，2021年，第 571–575 页。[16] K. Cho、B. Van Merri¨enboer、C. Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk 和 Y. Bengio，《使用 RNN 编码器-解码器进行统计机器翻译的短语表示学习》，*arXiv* 预印本 *arXiv:1406.1078*，2014年。[17] Y. Gong、C.-I. Lai、Y.-A. Chung 和 J. Glass，《SSAST：自监督音频频谱图变换器》，载于人工智能 *AAAI* 会议论文集，2022年，第 36 卷，第 10699–10709 页。[18] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weis- senborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，《一张图像胜过 16×16 个词：用于大规模图像识别的变换器》，*ICLR*，2021年。[19] H. Touvron、M. Cord、M. Douze、F. Massa、A. Sablay- rolles 和 H. J´egou，《通过注意力训练数据高效的图像变换器并进行蒸馏》，载于国际机器学习会议。PMLR，2021年，第 10347–10357 页。[20] Y. Wang、J. Li 和 F. Metze，《用于弱标签声音事件检测的五种多实例学习池化函数比较》，载于 *IEEEICASSP*，2019年，第 31–35 页。[21] N. Turpault、R. Serizel、J. Salamon 和 A. P. 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》，2019年。[22] C.. Bilen、G. Ferroni、F. Tuveri、J. Azcarreta 和 S. Krstulovi´c，《声音事件检测鲁棒评估框架》，载于 *IEEE ICASSP*，2020年，第 61–65 页。[23] Y. N. D. D. L.-P. Hongyi Zhang、Moustapha Cisse，《MixUp：超越经验风险最小化》，国际学习表征会议，2018年。[24] H. Nam、S.-H. Kim 和 Y.-H. Park，《FilterAugment：一种声学环境数据增强方法》，载于 *IEEE ICASSP*，2022年，第 4308–4312 页。[25] I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，*arXiv* 预印本 *arXiv:1711.05101*，2017年。