arXiv:2303.07626v1 [cs.SD] 14 Mar 2023

CAT：用于音频分类的因果音频变换器 N

Xiaoyu Liu[1], *Hanlin Lu*[2],*Jianbo Yuan*[2],*Xinyu Li*[2]

[1]马里兰大学帕克分校， [2]字节跳动

摘要

基于注意力的Transformer凭借其全局感受野和处理长期依赖关系的能力，越来越多地应用于音频分类。然而，现有框架主要由视觉Transformer扩展而来，与音频信号并不完全兼容。本文提出因果音频变换器（CAT），它包含一种多分辨率多特征（MRMF）特征提取方法，并结合声学注意力模块，以实现更优化的音频建模。此外，我们提出一种因果模块，可缓解过拟合、助力知识迁移并提高可解释性。CAT在ESC50、AudioSet和UrbanSound8K数据集上取得了更优或相当的最先进分类性能，且能够轻松推广至其他基于Transformer的模型。

索引词—音频分类、变换器、因果推断、注意力

I. 引言

音频分类的任务是分析音频录音，并为音频样本分配相应的标签。在Transformer[1] 在自然语言处理（NLP）和计算机视觉（CV）领域取得巨大成功之后，它们也被应用于音频模态，并取得了最先进的性能。这部分成功归功于Transformer中的全局感受野，使其能够捕获音频信号中的长距离上下文。现有的音频Transformer模型沿用了著名的视觉Transformer（ViT）的结构[2], , 主要是因为其最广泛使用的特征之一——梅尔频谱图，与图像格式相同。在图像建模中，x轴和y轴承载空间信息；而在音频输入中，梅尔频谱图的x轴表示时间信息，y轴则承载离散的频率信息 [3]–[8]。

尽管取得了优异的性能，这些热门的音频变换器仍面临尚未解决的挑战：（1）常用的声学表征采用不同的时频变换，包含不同尺度和粒度的声学语义 [4]–[6],，而视觉变换器 [2] 使用传统的自注意力、补丁采样和嵌入，难以有效捕捉这些语义；（2）在视觉任务中行之有效的设计，例如 ResNet [9],混合增强 [10], , 在声学领域更容易导致过拟合且泛化能力较差；（3）特征选择和表征学习在计算机视觉中至关重要，但在声学建模中却经常被忽视。

面对这些挑战，我们提出因果音频 Transformer（CAT），其包含多分辨率多滤波器（MRMF）特征提取、声学注意力和因果模块。频谱图是音频模型的标准输入，通过傅里叶变换采样，这自然导致时间分辨率与频率分辨率之间的权衡。CAT 通过提取不同分辨率和滤波器下全面的时频特征块来平衡这种权衡，随后将其与三维位置嵌入相结合。接着，我们提出声学注意力，以特征块为输入，有效地从此类表示中提取语义。不同滤波器产生的特征块均匀分布在各注意力头中；同时，我们计算不同分辨率、但处于相同时间帧内的特征块之间的成对注意力，从而实现不同粒度的信息交换。我们进一步引入因果模块，基于反事实推理 [11], [12]，建立学习到的表示与预测标签之间必要且充分的关系。我们将 [12] 拓展至音频分类领域，并给出了基于干预分布的必要性与充分性概率（PNS）下界。由于该下界只能从真实分布中估计，我们提出一种因果模块，学习从干预分布到观测数据集（即我们拥有的数据集）的映射，以缓解过拟合、提高可解释性并促进更好的知识迁移。

CAT 在 ESC50 [13], Au- dioSet [14]和 UrbenSound8K [15] 上达到了最先进性能。总而言之，我们的贡献如下：

- 因果音频变换器（CAT），采用多分辨率多滤波器（*MRMF*）特征提取和声学注意力进行声学建模。
- 采用重建模块的因果损失通过必要性与充分性概率（PNS）明确衡量特征质量，从而缓解过拟合，并提升不同数据集之间的知识迁移效果。

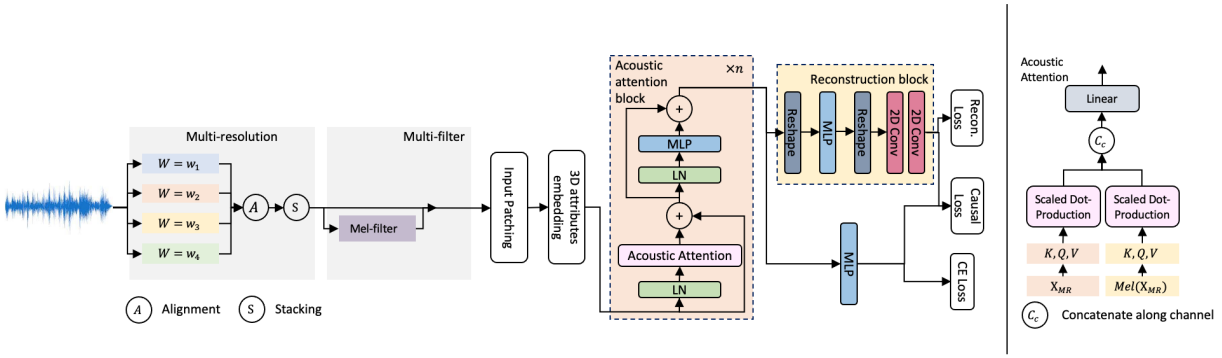

图1：左：所提出的CAT概览；右：所提出的声学注意力的详细结构。

II. 相关工作

音频分类：从卷积神经网络到变换器。音频分类是一项预测音轨标签的任务。大规模训练的巨大成功源于卷积神经网络模型在音频分类领域的兴起 [8], [16]–[20]。近年来，随着注意力机制在自然语言处理（NLP）和计算机视觉（CV）领域大获成功，AST 首次将基于自注意力的 Transformer 应用于声学分类 [5], ，并使用频谱图作为视觉变换器（ViT） [2]骨干网络的输入。之后的多项研究致力于提升模型效率：PaSST [4]提出了一种 patch-out 机制，而 HTS-AT [6] 采用了层次化骨干网络 [21]。为进一步提升性能，MBT [3] 引入了视觉-声学融合，而 PLSA [7] 则提出了一种与模型无关的框架。然而，近期大多数研究中的网络结构都大量借鉴了最初为自然语言处理和计算机视觉任务提出的原始变换器，因此更容易过拟合，对声学数据的泛化能力也较弱。因此，本文提出了融入专用于声学建模的 MRMF 特征的声学注意力。所提出的声学注意力可以轻松扩展至各种变换器模块。

特征表示中的因果推断。因果关系的概念最初是在图形概率模型中提出的 [11]。尽管因果推断在音频分类中是一个相对较新的概念，但它已推动可解释机器学习和表征学习取得进展。对生成因素之间的因果关系进行建模，能够显著促进代表性特征的学习 [22]。[23] 采用反事实信息，帮助在不同领域之间进行知识迁移。 [12]证明，学习到的表示是标签预测的必要且充分条件，其下界得以确立。然而，这一下界是在干预设置中得到的，因此，在未知真实分布的情况下无法直接估计。我们的设计基于 [12] ，并将其扩展为一个因果模块，在此背景下将下界映射到观测数据集音频分类。

III. 因果音频变换器

本节介绍我们的因果音频 Transformer（CAT）。CAT 首先利用三维位置嵌入提取 MRMF特征块。随后将这些特征块作为声学注意力的输入（第 III-A 节）。接着，我们提出一个因果模块，以缓解过拟合，并提升可解释性和知识迁移能力（第 III-B 节）。

III-A. 声学建模的变换器

图像建模与声学建模大不相同，因为二维图像仅沿其轴携带空间信息，而声学特征则包含时间信息和频率信息。因此，直接应用传统的视觉Transformer来实现高效的声学建模并非易事。我们通过提取MRMF特征并采用声学注意力来弥合这一差距。

多分辨率多滤波器特征提取。由于一维傅里叶变换（FT）的性质，在时频表示（例如频谱图）中，频率分辨率与时间分辨率之间存在权衡。我们提出提取具有不同时间分辨率的频谱图，如下所示：

$$x_{\mathrm{MR}} = \{\mathrm{FFT}_{\omega}(s)\}, \omega \in \{\omega_1, \omega_2, ..., \omega_K\} \tag{1}$$

其中，s 表示输入音频波，FFT_{ω} 表示窗长为 ω 的一维快速傅里叶变换，$\{*\}$ 表示一个集合。对于每个原始输入 x，x_{MR} 包含 k 个频谱图，每个频谱图的大小为 $\mathrm{R}^{T_i \times F}$，$i = 1, 2, ..., K$。$T_i$ 表示时间帧数，F 表示频率分箱数。这类多分辨率频谱图与 [24], [25], 具有相似的直觉，但我们的方法结合了声学注意力和三维位置嵌入（将在下一节介绍），因此能够更有效地利用信息。我们还允许采用多个滤波器的频谱图作为不同注意力头的并行输入。例如，我们可以将 x_{MR} 的副本输入 Mel 滤波器组，从而得到如下 MRMF 特征 $x_{\mathrm{MRMF}} \in \mathbb{R}^{\{T_i\} \times K \times F \times 2}$ ：

$$x_{\mathrm{MRMF}} = \{\mathrm{mel}(x_{\mathrm{MR}}), x_{\mathrm{MR}}\} \tag{2}$$

不同分辨率的频谱图大小各不相同。对齐它们有多种方法。一种可能的方法是沿时间维度拼接它们，形成超级频谱图。另一种方法是将频谱图投影到相同的时间维度。根据实证结果，我们采用后一种设置。

分块与三维位置嵌入。输入x_{MRMF} 被划分为特征块，并通过窗长 $Wi \in \mathrm{R}^{1\times K\times F\times 1}$进行聚合，我们使用线性投影 $\{\xi\}$ 将维度匹配至 M。与计算机视觉中方形输入的分块方式不同 [2], ，所提出的特征块x_p 包含不同分辨率的频率信息。这种设计聚合并保留了重要的多尺度频率信息，这对音频分类至关重要。受视频变换器中所用三维位置嵌入的启发 [26], 我们提出了一种声学三维位置嵌入，使网络能够感知使用不同窗长和滤波器提取的特征：

$$pe = (g([pe_1, pe_3]) \otimes \mathbf{1}_F)^T + pe_2 \otimes \mathbf{1}_T \tag{3}$$

其中， g 是线性投影， pe_1 是 T 维正弦时间嵌入， pe_2 是 F 维频率正弦嵌入，遵循 [1]。 pe_3 是分辨率的独热编码。 $\mathbf{1}_F$ 是 F维全一向量。

声学注意力。因此，我们设计了声学注意力，以便在音频变换器中纳入 MRMF 特征块。具体而言，多滤波输入被送入不同的注意力头，并分别通过缩放点积运算进行处理（图1），从而独立计算经过梅尔滤波和未经梅尔滤波的注意力。需要指出的是，所提出的声学注意力适用于各种注意力核（例如多头注意力 [1], SWIN注意力 [21],等）。我们的实验采用 SWIN 核。

III-B. 用于更好表示的因果模块

在 [12], 中首次提出了这样一种思路：通过最大化干预下的下界，促使学习到的隐藏表示成为标签预测的必要且充分原因。 [12]证明，这有助于构建过拟合更少、鲁棒性更强的模型。然而，这种下界需要估计未观测混杂因素C, 因此难以在实证设置中取得成功。本文引入一个带有因果损失的因果模块，通过对双射映射施加额外约束来解决这一问题。本节首先从因果视角介绍预备知识，然后介绍所提出的模块。

预备知识 令 X 表示训练数据， Y 表示反事实标签函数。我们定义函数 f ，将 X 映射为其潜在表示 Z，其中 $Z = f(X)$。

$\mathbb{I}\{Z = z\}$对于 $\mathbb{I}\{Y = y\}$ 的充分性与必要性概率定义为 [12]:

$$\text{PNS}_{Z=z,Y=y} = P(Y(Z \neq z) \neq y, Y(Z = z) = y). \tag{4}$$

其中， Y（$Z = z$）推断如果 $Z = z$，标签会是什么。为防止过拟合并确保特征具有代表性，我们的目标是在训练期间最大化 PNS。 [12]建立了 PNS 的下界，如下所示：

$$\text{PNS} \geq P(Y = y|do(Z = z)) - P(Y \neq y|do(Z \neq z)) \tag{5}$$

其中， do 符号 [11] 表示对 Z进行干预。

然而，在不知道真实分布的情况下，无法直接估计对潜在表示 Z 进行特定干预后相应的标签变化。我们不通过去混杂来解决这一问题，而是推导出下界的一个估计，如下所示：

$$\text{PNS} \geq= \int P(Y|X)[P(X|f(X) = z) - P(X|f(X) \neq z)]dX \tag{6}$$

从概率视角来看，我们可以观察到，最大化$P(Y|X)$即为分类目标，而 $P(X|f(X) = z)$则是根据特定潜在表示 $Z = z$ 推断出 X 的概率。鉴于我们希望最大化 PNS 的这一类下界，一种可行的策略是同时最大化 $P(Y|X)$和 $P(X|f(X) = z) - P(X|f(X) \neq z)$。前者符合分类目标，而第二项则表明，我们需要在 Z 与 X之间实现准确的双射匹配。

重构与因果损失 我们认为，可以通过采用重构模块和“因果”损失来实现这种映射，如图所示1。我们提出了一个简单的重建模块，与分类层并行运行，使得 X 和 Z 之间的近似双射关系可通过重构损失 l_{rs}进行度量：

$$l_{rs} = \|\Phi(Z) - X\|_2 \tag{7}$$

其中 Φ 由图1所示的重建模块进行参数化。

在满足上述要求的前提下，我们进一步最小化因果损失（ [12] 公式 25）：

$$lc = -\sum_{j=1}^{d} log\text{PNS}_n(f_j(X), Y|f_{-j}(X)) \tag{8}$$

其中

$$\text{PNS}(f_j(X), Y|f_{-j}(X)) = \prod_{i=1}^{n} \text{PNS}_{f_j(X_i),Y_i|f_{-j}(X_i)} \tag{9}$$

CAT 的目标是交叉熵损失l_θ、重构损失和因果损失之和：

$$L = \text{argmin}(l_\theta + l_c + l_{rs}) \tag{10}$$

IV. 实验

IV-A. 数据集

我们在三个数据集上评估 CAT。**AudioSet [14]**：包含超过两百万个带有 527 个标签的 10秒音频片段。我们按照 [5]报告评估集上的平均精度均值（mAP）。**环境声音分类数据集 50 [13]:** 包含 50 个类别和 2000 条音频录音。我们按照 [5]报告 5 折交叉验证的平均准确率和 5 轮实验的 mAP。**UrbanSound8K [15]:** 包含 8732 个带标签的声音片段，涵盖 10 个类别。我们按照 [27]报告十折交叉验证的平均准确率。

IV-B. 实现细节

遵循先前的方案 [4], [5], [8], , 所有实验均使用 32kHz 作为采样率。我们以 25ms 的窗长和 10ms 的帧移生成频谱图和梅尔频谱图。训练时，我们采用反比例加权 [28]来平衡 AudioSet训练数据，并在所有实验中使用 0.5 [29] 的混合增强。

IV-C. 主要结果

表 I：在 *AudioSet* 和 *ESC50* 数据集上的评估。我们报告 *AudioSet* 的 *mAP* 和 *ESC50* 的 *Top-1*平均准确率。* 表示该模型的集成版本。

模型	模态	AudioSet %wo	ESC50 %
PANN [8]	音频	0.434	94.7
AST [5]	音频	0.347	95.6
HTS-AT [6]	音频	0.471	-
CAT	音频	**0.478**	**96.4**
PSLA* [7]	音频	0.474	-
AST* [5]	音频	0.475	95.7
HTS-AT* [6]	音频	0.487	97.0
PaSST* [4]	音频	0.496	96.8
CLAP [30]	音频+文本	-	96.7
AnT [31]	音频+视频	-	95.7
CAT*	音频	**0.489**	**96.9**
CAT + PANN*	音频	**0.491**	**97.2**

AudioSet 我们在表 I中展示了 CAT 在 AudioSet 上的表现。我们的模型显著优于以卷积神经网络为基础的既有工作 [8], , 也以明显优势超过了基于变换器的最先进模型（+0.7%相比于[6]）。值得注意的是，我们的 CAT 与 HTS-AT 采用了相似的 SWIN 骨干网络，其优异表现证明了我们提出的、结合 MRMF 和声学注意力的音频 Transformer，以及因果模块的有效性。值得一提的是，遵循相同的集成协议，CAT 的表现可与最先进的集成模型相媲美。只需将 CAT 与 PANN 集成 [8],，我们便实现了与此前 9 模型集成 [4]结果相当的表现。这也表明，CAT 能为基于卷积神经网络的模型提供互补信息。

ESC50 CAT 在 ESC50 数据集上也达到了最先进性能（表 I）。我们观察到类似的模式，即

表 II：在 *ESC50* 和 *AudioSet* 上进行的消融实验。*CL* 表示因果损失，*US8K* 表示 *UrbanSound8K*。

（a）因果损失有助于提升两个数据集上的表现。

	AudioSet	*ESC50*
无对比学习	0.474	95.9
采用对比学习	0.478	96.4

（b）预训练有助于提升两个数据集上的性能。

预训练	*AudioSet*	*ESC50*
N/A	0.451	89.4
ImageNet	0.478	96.4

(c) MRMF 和对比学习能够泛化到不同的骨干网络。

	AST [5]	*CAT*
-	88.6	88.7
+ MRMF	88.9	89.1
+ CL	89.1	89.4

(d) 所提出的因果损失有助于特征域适应。

预训练	*val*	不使用对比学习	使用对比学习
US8K	ESC50	84.1	86.7
ESC50	US8K	95.9	96.1

所提出的 CAT 在使用和不使用集成时，分别比此前的最先进水平 [5], [6] 高出 0.8% 和 0.2%，证明了所提设计的有效性。值得注意的是，将 CAT 与其他声学模型 [8]集成后，其性能可媲美此前使用额外模态训练的工作。例如，CAT 的表现比利用视觉-声学信息的 AnT [30] 高出 0.7%。可以预期，将多模态特征融入 CAT 可能会进一步提升性能。

IV-D. 消融实验

我们在 ESC50 和 AudioSet 上进行了消融实验。**因果损失的有效性**。表IIa显示，因果损失有助于提升不同规模数据集上的性能。

预训练的影响。ImageNet 预训练有助于提升两个数据集上的性能（表IIb），但其在 ESC50 上的影响更为显著。这是因为变换器通常需要大规模数据进行训练。

泛化能力我们表明，所提出的 MRMF 和因果损失能够很好地泛化到不同的骨干网络（例如，采用 [21] 和未采用 [5] 层级结构的骨干网络），并在表IIc中持续提升性能。

声学嵌入质量 我们证明，所提出的因果损失能够消除噪声，从而提升跨数据集的特征迁移效果（表 IId）。我们在小规模数据集上进行预训练，以确保性能提升源于所提出的因果损失，而非数据重叠。

V. 结论与未来工作

本文提出了因果音频变换器（Causal Audio Transformer），这是一种专为音频分类设计的声学变换器。CAT采用MRMF特征、声学注意力和因果模块。我们通过实验证明，所提出的CAT在多个数据集上达到了最先进性能。消融实验还表明，MRMF和因果损失能够很好地泛化到不同的骨干网络。

VI. 参考文献

[1] Ashish Vaswani等人，“注意力就是一切”，神经信息处理系统进展，第30卷，2017年。[2] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly等人，“一张图像胜过16×16个词：用于大规模图像识别的变换器”，*arXiv*预印本 *arXiv:2010 11929*，2020年。[3]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，第14200–14213页，2021年。[4] Khaled Koutini、Jan Schl¨uter、Hamid Eghbal-zadeh和Ger- hard Widmer，“使用补丁丢弃高效训练音频变换器”，*arXiv*预印本 *arXiv:2110.05069*，2021年。[5] Yuan Gong、Yu-An Chung和James Glass，“AST：音频频谱图变换器”，*arXiv*预印本 *arXiv:2104 01778*，2021年。[6] Ke Chen等人，“HTS-AT：用于声音分类与检测的分层令牌语义音频变换器”，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第646–650页。[7] Yuan Gong、Yu-An Chung和James Glass，“PSLA：通过预训练、采样、标注和聚合改进音频标签”，*IEEE/ACM*音频、语音与语言处理汇刊，第29卷，第3292–3306页，2021年。[8] Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang和Mark D Plumbley，“PANN：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。[9] Alex Krizhevsky、Geoffrey Hinton等人，“从微小图像中学习多层特征”，2009年。[10] Hongyi Zhang、Moustapha Cisse、Yann N Dauphin和David Lopez-Paz，“混合增强：超越经验风险最小化”，*arXiv*预印本 *arXiv:1710.09412*，2017年。[11] Judea Pearl，《因果关系》，剑桥大学出版社，2009年。[12] Yixin Wang和Michael I Jordan，“表征学习的期望：因果视角”，*arXiv*预印本 *arXiv:2109.03795*，2021年。[13] Karol J Piczak，“ESC：环境声音分类数据集”，载于第*23*届*ACM*国际多媒体会议论文集，2015年，第1015–1018页。[14] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal和Marvin Ritter，“Audio Set：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。[15] Justin Salamon、Christopher Jacoby和Juan Pablo Bello，“用于城市声音研究的数据集与分类体系”，载于第*22*届*ACM*国际多媒体会议论文集，2014年，第1041–1044页。[16] Paulo Lopez-Meyer、Juan A del Hoyo Ontiveros、Hong Lu和Georg Stemmer，“面向目标应用音频分类的高效端到端音频嵌入生成”，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第601–605页。[17] Jaehun Kim，“使用卷积神经网络和多通道音频特征进行城市声音标签”，论文集

声学场景与事件检测和分类，2020年。[18]Loris Nanni、Gianluca Maguolo、Sheryl Brahnam 和 Michelangelo Paci，“用于音频分类的卷积神经网络集成”，*AppliedSciences*，第 11 卷，第 13 期，第 5796 页，2021年。[19]Hardik B Sailor、Dharmesh M Agrawal 和 Hemant A Patil，“使用卷积受限玻尔兹曼机进行环境声音分类的无监督滤波器组学习”，载于 *InterSpeech*，2017年，第 8 卷，第 9 页。[20] Jonathan J Huang 和 Juan Jose Alvarado Leanos，“Aclnet：高效的端到端音频分类卷积神经网络”，*arXiv* 预印本 *arXiv:1811.06669*，2018年。[21] Ze Liu 等人，“Swin Transformer：使用移位窗口的层次化视觉变换器”，载于*IEEE/CVF*国际计算机视觉大会论文集，2021年，第 10012–10022 页。[22] Mengyue Yang、Furui Liu、Zhitang Chen、Xinwei Shen、Jianye Hao 和 Jun Wang，“CausalVAE：通过神经结构因果模型进行解耦表示学习”，载于 *IEEE/CVF*计算机视觉与模式识别大会论文集，2021年，第 9593–9602 页。[23] Sharmila Reddy Nangi、Niyati Chhaya、Sopan Khosla、Nikhil Kaushik 和 Harshit Nyati，“利用反事实控制用于风格迁移的潜在解耦文本表示”，载于计算语言学协会第 *59* 届年会暨自然语言处理国际联合会议第 *11* 届会议论文集（第 *2* 卷：短论文），2021年，第 40–48 页。[24] Boqing Zhu、Changjian Wang、Feng Liu、Jin Lei、Zhen Huang、Yuxing Peng 和 Fei Li，“使用多尺度卷积神经网络学习环境声音”，载于 *2018*年国际神经网络联合会议（*IJCNN*）。IEEE，2018年，第 1–8 页。[25] Emad M Grais、Hagen Wierstorf、Dominic Ward 和 Mark D Plumbley，“用于单声道音频源分离的多分辨率全卷积神经网络”，载于潜变量分析与信号分离国际会议。施普林格，2018年，第 340–350 页。[26] Yanyi Zhang 等人，“VidTr：无需卷积的视频变换器”，载于*IEEE/CVF*国际计算机视觉大会论文集，2021年，第 13577–13587 页。[27] Avi Gazneli、Gadi Zimerman、Tal Ridnik、Gilad Sharir 和 Asaf Noy，“端到端音频卷土重来：增强数据增广以构建高效音频分类网络”，*arXiv* 预印本 *arXiv:2204.11479*，2022年。[28] Pavel Izmailov、Dmitrii Podoprikhin、Timur Garipov、Dmitry Vetrov 和 Andrew Gordon Wilson，“权重平均带来更宽的最优解和更好的泛化能力”，*arXiv* 预印本 *arXiv:1803.05407*，2018年。[29]Yuji Tokozume、Yoshitaka Ushiku 和 Tatsuya Harada，“利用类间样本进行深度声音识别”，*arXiv* 预印本 *arXiv:1711.10282*，2017年。[30]Benjamin Elizalde、Soham Deshmukh、Mahmoud Al Ismail 和 Huaming Wang，“CLAP：通过自然语言监督学习音频概念”，*arXiv* 预印本 *arXiv:2206.04769*，2022年。[31] Yanpeng Zhao、Jack Hessel、Youngjae Yu、Ximing Lu、Rowan Zellers 和 Yejin Choi，“通过视觉知识迁移，在没有平行数据的情况下建立音频与文本之间的联系”，*arXiv* 预印本 *arXiv:2112.08995*，2021年。