arXiv:2305.01578v1 [cs.SD] 2 May 2023

用于婴儿哭声分析的自监督学习

Arsenii Gorin[⋆]、*Cem Subakan*[♮♯♭]、*Sajjad Abdoli*[⋆]、*Junhao Wang*[⋆]、*Samantha Latremouille*[⋆]、*Charles Onu*[⋆♭]

[⋆]Ubenwa Health，Montr´eal，加拿大 [♮]拉瓦尔大学，魁北克市，加拿大[♯]康考迪亚大学，Montr´eal，加拿大 [♭]Mila-Qu´ebec 人工智能研究所，Montr´eal，加拿大

摘要

本文探讨了自监督学习（SSL）在分析一个开创性哭声录音数据库中的应用。该数据库包含一千多名新生儿的临床指征。具体而言，我们关注基于哭声检测神经系统损伤，以及识别疼痛、饥饿和不适等哭声诱因。在医疗环境中，为大型数据库添加标注既昂贵又耗时，通常需要多位专家多年协作。利用大量未标注音频数据学习有用的表征，可以降低构建稳健模型乃至临床解决方案的成本。在这项工作中，我们尝试在大型音频数据集上对卷积神经网络进行自监督预训练。结果表明，对于神经损伤检测和哭声诱因识别，采用自监督学习对比损失（SimCLR）进行预训练，效果显著优于监督式预训练。此外，我们还展示了使用未标注婴儿哭声进行基于自监督学习的领域自适应，可以进一步提升性能。我们还发现，使用这类基于自监督学习的预训练来适应哭声，可减少整个系统对标注数据的需求。

索引术语— 自监督学习，婴儿 cry 分类，音频分类，迁移学习，领域自适应

1. 引言

哭泣是婴儿与外界交流的主要方式。自 20 世纪 60 年代初以来，研究人员一直对婴儿哭声分析感兴趣 [1]。哭声特征可能有助于我们理解婴儿的基本需求（饥饿、疼痛等），更重要的是，还可以通过分析这些特征，及早、无创地检测各种疾病 [2]。例如，临床研究报告称，某些婴儿哭声特征与出生窒息相关 [3]。这种由多种因素引起的状况常常会导致严重的健康问题，包括神经系统损伤，甚至死亡。研究人员已探索多种基于信号处理 [4], 、统计建模 [5, 6] 和深度学习 [7–10]的方法，以利用哭声录音发现临床及其他方面的见解。

婴儿哭声分析面临的主要挑战之一是数据获取。目前，哭声并非常规医疗记录的一部分，因此要建立数据库，需要开展临床研究等有针对性的工作。这类研究成本高昂，通常需要多名医院工作人员多年协作完成。大多数利用哭声进行病理检测的机器学习（机器学习）研究都使用了 Baby Chillanto [11] 数据库，该数据库仅包含六名确诊为出生窒息的患者。

从机器学习问题的角度来看，哭声分类类似于通用音频分类，而深度卷积神经网络（CNN）在这一领域已达到最先进水平。近期， [12] 证明，预训练音频神经网络（PANNs）——在通用音频数据上预训练的大型卷积神经网络——迁移到广泛的音频模式识别任务后，表现优于此前若干最先进的系统。此后，PANNs 已广泛应用于各种音频任务，包括从语音中识别情绪 [13] ，以及通过咳嗽检测 COVID-19 [14]。

音频分类领域当前最先进方法中的另一种流行范式是自监督学习（SSL）——通过在无标签数据上训练来获得高质量表征的方法。自监督学习（SSL）彻底改变了自然语言处理和计算机视觉领域，目前已广泛应用于音频处理 [15]。通过自监督学习（SSL）预训练的神经网络（编码器），可以视为将音频序列映射到隐藏表示（一种嵌入表示）的非线性映射。嵌入表示可作为分类器的输入，该分类器针对特定任务，使用监督目标进行训练（使用标注数据和传统的交叉熵损失）。这种方法常用于在多种不同的音频任务上对各种自监督学习（SSL）模型进行基准测试 [16, 17]。近期，一种名为 SimCLR、基于相似度的对比学习方法在计算机视觉领域被提出 [18, 19]，并在多种音频任务中展现出良好表现 [17, 20], , 其中包括音乐分析 [21, 22]。SimCLR 会最大化同一对象不同变换（失真）视图之间的相似度。对于音频，可以通过混合随机音频样本 [17], 、频谱图掩蔽 [23] 、 [20], 中的混响、音高移位等方式进行此类变换 [21]。

本文中，我们使用监督式预训练和自监督预训练的 PANNs 进行实验，为两个下游任务学习表征。第一个任务是对脑损伤（由出生窒息导致）进行分类，第二个任务是预测哭声诱因（疼痛、饥饿、不适）。我们在一个独特的临床数据库上测试了这些方法，该数据库收录了 Ubenwa Health 与三个国家的医院合作采集的新生儿哭声 [24]。

此外，我们还借鉴了语音 [25] 和自然语言处理 [26]领域中的自监督域适应方法，利用未标注的啼叫声，评估了基于 SimCLR 的 PANNs 适应方法所产生的影响。值得注意的是，语音和音频领域的自监督学习（SSL）最先进方法通常采用变换器而非卷积神经网络，并依赖不同的学习目标 [15]。然而，我们使用一些流行的预训练语音和音频变换器开展的初步实验（具体包括 Wav2Vec2.0 [27], HuBERT [28], WavLM [29] 和 SSAST [30]）并未取得足够显著的性能提升，而且通常需要适配大量参数并调整超参数。因此，我们将重点放在卷积神经网络和 SimCLR 上，因为二者在准确率与适应复杂度之间实现了良好的平衡。

2. 方法

我们在图 1 中展示了三阶段方法，并标明了各阶段使用的自监督学习流程和数据集。本节其余部分将详细介绍各个阶段。

© 2023 IEEE。允许个人使用本材料。对于任何其他用途，无论是在当前还是未来的媒体中，包括为广告或促销目的重印或重新发布本材料、创建新的合集、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护部分，均须事先获得 IEEE 的许可。

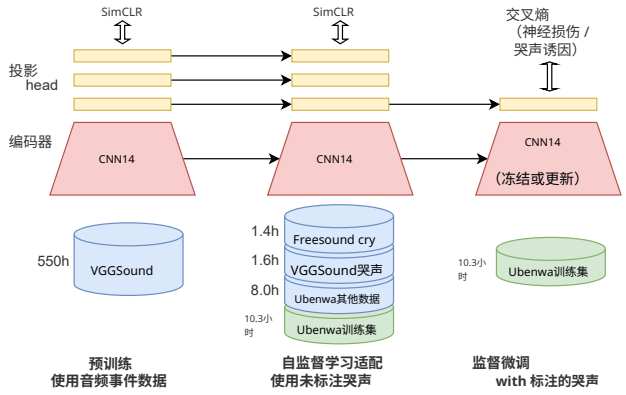


图1：所提出的基于自监督学习的训练流程概述。**（左）**首先，使用 SimCLR 在 VGGSound 数据集上对 CNN14 主干网络进行预训练。**（中）**随后，使用哭声专用数据集通过自监督学习进一步预训练 CNN14 主干网络。我们将这一阶段称为*SSL*哭声适应。**（右）**最后，使用带标签的哭声数据集对模型进行监督训练。

2.1. 编码器架构

我们采用 [12] 中介绍的 CNN14 作为骨干编码器。该编码器按照 [20] 中的方法，在 VGGSound 数据库上进行预训练。

[12] 中提出的 CNN14 与我们从 [20]采用的模型之间存在一些细微差异。首先，[20] 使用 VGGSound [31] 而非 AudioSet [32] 进行预训练。VGGSound 是一个通过自动整理而成的 YouTube 录音集合，时长为 550 小时，约为 AudioSet 的十分之一，但每段音频录音只有一个标签，且噪声水平较低。其次，[20] 在训练中使用了更多的对数梅尔滤波器组（80 个而非 64 个），以及更短的音频片段（4 秒而非 10 秒）。整个网络约有 8000 万个参数。

2.2. 预训练

我们比较了两个完全相同的 CNN14 模型，它们分别以监督学习和自监督学习的方式在 VGGSound 上进行预训练。SimCLR 预训练依赖于一个额外的投影头——一个包含三层的多层感知机，其隐藏单元数为 2048，并带有一个包含 512 个隐藏单元的瓶颈层。在监督学习和自监督学习的初始预训练过程中，模型均采用随机梯度下降进行更新，每个小批次包含 32 个四秒片段，这些片段从 VGGSound 音频片段中随机截取。在提取验证集和测试集数据的嵌入表示时，模型使用全局时间池化处理任意长度的音频序列。

2.3. 监督微调与评估

为研究预训练模型的各个方面，我们进行了三项评估。首先，我们采用线性探测，由冻结的编码器提取特征，供线性分类器使用。其次，除了学习线性分类器外，我们还更新编码器批量归一化参数的统计量，同时冻结其他参数。这样做的主要目的是弥补预训练数据与目标数据特征之间的差异 [33, 34]。这也使我们能够了解，在使用 cry 数据进行自监督学习微调所获得的提升中，有多少可能归因于 SimCLR 过程中自然发生的归一化参数简单更新。最后，为进一步提升分类结果，我们进行端到端微调，在目标任务上联合优化编码器参数和分类器参数。

在所有三种设置中，监督训练均进行 50 个训练轮次，并选择验证分数最佳的模型。训练期间，我们使用加权随机采样来平衡各类别的分布。我们使用 Adam 优化器 [35]，如果验证损失连续三个训练轮次未改善，则将学习率降低两次。对于端到端监督微调，编码器使用小得多且单独调优的学习率。此外，编码器的学习率在前十个训练轮次中从零线性递增至目标值。在所有实验中，分类器和编码器的学习率均通过网格搜索进行优化。对于验证分数最佳的模型，我们重复实验 10 次，并报告受试者工作特征曲线下面积（曲线下面积）的均值及标准误。对于多类分类（触发器），采用一对多方法计算宏平均曲线下面积。

与我们在初步研究中的发现类似，保留 SSL 预训练模型的一个投影头层会带来略好的结果。因此，我们始终从投影头的第 1 层进行迁移。

2.4. 自监督哭声适应

为提高自监督学习（SSL）表征的质量，我们进一步探索了第二阶段的自监督域适应。我们的目标是利用未标注的哭声数据，将编码器从通用音频声音领域适配到哭声领域（图 1 中间列）。这一自监督学习适配通过复用在 VGGSound 上训练的 CNN14 的编码器和投影器实现，并使用婴儿哭声数据进一步运行 100 个训练轮次的 SimCLR，其学习率和训练计划与初始预训练阶段相同。唯一的区别在于，为了加快训练速度，并且考虑到已有文献表明，较大的批量大小在 SimCLR 中表现更好 [21]，我们采用了 200 的批量大小。值得注意的是，我们发现，在不使用哭声音频的情况下，通过增加训练轮次和采用较大的批量大小来改进 [20] 的初始自监督学习预训练模型，并未带来显著差异。

与初始的自监督学习预训练模型类似，经过 cry 适配的模型也采用上一节所述的线性探测和端到端设置进行评估。

3. 实验设置

3.1. 数据集描述

本研究基于 Ubenwa 新生儿哭声临床数据库的一个子集。该数据库规模更大，自 2020 [24]以来在尼日利亚、巴西和加拿大的五家医院采集。大多数婴儿在出生后和出院前各录制一次。所有婴儿均接受了神经系统检查，并使用名为 Sarnat评分 [36]的四级量表记录神经损伤程度。进行分类时，我们将录音分为两组：正常（无神经损伤）和受损（轻度、中度或重度损伤）。我们进一步将数据划分为训练集、验证集和测试集，并确保同一患者的录音只分配至其中一个子集。表1总结了数据的关键统计信息。

此外，这些录音还标注了一个诱因，即由医务人员或研究人员确定的哭泣首要原因。本研究使用了三种主要诱因的一个子集，包含267段不适录音、200段饥饿录音和682段疼痛录音。

	健康			神经损伤		
	训练	Val	Test	训练	Val	Test
# 录音	1360	247	238	92	40	45
# 患者	885	165	163	75	33	38
# 小时	10.3	1.9	2.0	0.8	0.3	0.3

表1：神经系统损伤数据集的描述。

与常用的Chillanto数据库 [11], 相比，我们的数据集中有更多神经系统损伤患者（146名，而非6名），总体上也包含更多已标注的哭声信号（14.2小时，而非0.6小时）。在我们的数据库中，一段哭声录音的时长不固定，从一秒到几分钟不等。相比之下，Chillanto中的录音对应于1秒的哭声呼气段，并标注为来自健康或患病婴儿。然而，现有证据不足以确定患病婴儿的每个哭声呼气段是否都具有区别于健康婴儿的特征，还是只有部分呼气段具有这些特征。此外，同一婴儿的哭声呼气段通常非常相似，因此，如果在训练、测试集和验证集中随机划分录音，而不考虑婴儿身份（例如，在 [10]中），最终的性能可能会被高估。

对于自监督学习实验，我们还使用了额外的 8 小时 Uben wa 未标注哭声，以及 VG-GSound 中可用的 1.6 小时和从 Freesound 网站[1]通过搜索查询“baby cry”收集的约 1.4 小时哭声。

3.2. 基线方法

虽然我们的主要重点是自监督学习预训练和微调，但我们采用两种不依赖预训练的监督方法作为基线。第一个系统是一个统计模型，使用通过 OpenSmile 工具包 [38]提取的 ComParE 2016 [37] 声学特征。该特征集包含 6373 个录音级衍生特征（均值、标准差等），它们来自各种声学描述符（梅尔频率倒谱系数、基频、微扰等）。这些特征常用于计算副语言学和婴儿哭声分类 [6]。模型及超参数通过网格搜索和 10 折交叉验证进行选择，以平均曲线下面积得分最大化为目标。本研究在模型选择时考察了随机森林、逻辑回归和支持向量机。第二个基线是第 2 节所述的 CNN14，采用随机初始化，不进行预训练，并按照第 2.3 节所述进行端到端监督微调。不同随机种子下的五组实验中，基线方法在神经系统损伤和诱因上的表现汇总于表2。未经预训练的 CNN14 表现不及统计基线；鉴于我们的训练数据集对于此类模型而言相当小，这一结果并不令人意外。

模型	**曲线下面积 %（均值和标准误）** 神经损伤	哭声诱因
统计学的	75.1 ± 0.4	71.1 ± 0.2
CNN14	74.6 ± 1.7	59.8 ± 0.9

表 2：未进行任何预训练时获得的基线性能

4. 自监督学习实验

神经系统损伤和哭声诱因实验的主要结果汇总于表 3 和表 4 中。两个表格的第一行均指

[1]https://freesound.org

采用随机初始化进行监督训练，旨在展示模型未经预训练时的性能。

表3和表4中的最后三列总结了采用监督训练对网络进行微调后获得的性能。从左到右，表中的结果分别对应于以下内容:

1. 使用线性探测进行评估，即在线性层上对冻结的编码器权重进行微调（线性）。2. 使用线性探测进行评估，并在微调期间更新批归一化层（线性+批归一化）。3. 进行端到端微调，同时更新线性层和整个编码器（端到端）。

	Pre-训练	自监督学习 cry 适配。 **数据集**	**微调后的曲线下面积百分比** 线性	线性+批归一化	端到端
1	–	–	60.6 ± 1.5	60.4 ± 1.3	74.6 ± 1.7
2	监督学习	–	75.5 ± 0.6	75.9 ± 0.7	80.0 ± 0.7
3	SSL	–	71.3 ± 0.9	78.5 ± 1.2	83.9 ± 0.6
4	SSL	训练集	78.8 ± 0.5	78.0 ± 0.7	80.8 ± 0.8
5	SSL	+ 1 lh cry	79.8 ± 0.4	81.3 ± 0.5	81.3 ± 0.7
6	SSL	+ replayVGG	80.8 ± 0.5	83.3 ± 0.6	**85。0** ± 0.9

表3：不同类型的预训练和微调策略下神经损伤分类的表现。**第 1 列**表示预训练类型。请注意，第一行未应用预训练；第二行应用了基于 VG-GSound 数据集的监督式预训练。其余各行则使用基于 VGGSound 的自监督学习预训练。**第 2 列**表示自监督学习哭声适应所用的数据集。请注意，第 4、5、6 行中用于自监督学习哭声适应的数据集是逐步累积的。第 4 行使用神经损伤训练集；第 5 行在神经损伤训练集的基础上加入 11 小时哭声；第 6 行则在第 4、5 行所用的数据集基础上，加入来自 VGG Sound 数据集的重放缓冲区。**第 3–5 列**展示了采用不同监督微调策略（自监督学习微调之后，如图 1 所示）得到的 AUC 百分比（含均值和标准误）。

	Pre-训练	自监督学习 cry 适配。 **数据集**	**微调后的 % 曲线下面积** 线性	线性+批归一化	端到端
1	–	–	57.1 ± 2.6	60.7 ± 0.5	59.8 ± 0.9
2	监督学习	–	67.9 ± 1.9	68.0 ± 1.7	68.1 ± 1.6
3	自监督学习	–	65.9 ± 0.8	69.5 ± 0.7	69.0 ± 0.9
4	自监督学习	神经损伤训练集	71.7 ± 0.5	**75.4** ± 0.8	72.4 ± 1.4
5	自监督学习	+ 1 lh cry	74.5 ± 0.4	74.7 ± 0.4	72.0 ± 1.8
6	自监督学习	+ replayVGG	74.2 ± 0.4	**75.6** ± 0.6	74.4 ± 0.7

表4：哭声触发分类的性能。我们遵循表3所采用的相同结构，因此适用相同的标题。

两个表格中的第二行和第三行比较了使用 VGGSound 进行的监督式预训练和自监督初始预训练。在这些实验中，哭声数据库仅用于监督微调（换言之，未采用额外的 SSL 哭声适应阶段）。我们观察到，虽然简单的线性探测在监督式预训练中表现更好，但更新批归一化统计量时，自监督预训练的效果更佳。此外，我们还发现，采用端到端微调策略时，SSL 预训练在神经损伤任务上的表现显著更好。

接下来，在表3和表4的第4、5、6行中，我们展示了采用额外哭声自监督学习适配阶段时的结果。首先，我们

仅使用神经损伤训练数据集，通过 SimCLR 进行微调，如表3第4行所示）。对于神经损伤，我们观察到，这显著提高了线性评估的曲线下面积（从71.3提高到78.8），但在更新批归一化后，这一提升便消失了（78.5和78.0）。我们推测，在此实验中，SimCLR 表现更好，主要是因为 VGGSound 存在显著的领域失配，而简单的批归一化更新可以弥补这一点。对于哭声诱因，我们观察到，哭声自监督学习适配通常能提升监督微调后的表现。

接下来，如表3和表4第5行所示，我们进一步将 SimCLR 自监督学习 cry 适应阶段的无标签数据量翻倍。具体做法是加入此前未使用的 8 小时 Ubenwa哭声，以及来自 VGGSound 和 Freesound 的一些未标注哭声。这些自监督学习 cry 适应数据集的录音总时长约为 11 小时（因此在表3和表4中称为 11 小时哭声）。这显著提升了神经损伤任务中有无批归一化更新时线性评估的性能。

然而，端到端微调的情况并非如此：初始的 VGGSound 预训练使模型在神经损伤迁移方面表现更好（表3第3行）。我们假设，当我们使用自监督学习进一步调整模型时，由于灾难性遗忘，模型失去了有助于微调的泛化特性 [39] 。

为减轻这种遗忘效应，并保留对迁移学习似乎至关重要的泛化特性，我们采用持续学习文献中的回放技术进行哭声自监督学习适配 [40]。我们在两个表格的最后一行展示了这一结果。在哭声自监督学习适配阶段，我们通过回放 50% 的 VGGSound 数据集来实现这一点。因此，我们观察到，采用回放的哭声自监督学习适配在神经损伤的所有评估中表现都显著更好。采用这种方法，通过线性+批归一化评估进行触发器分类，也取得了最佳结果。

5. 使用标注数据子集进行训练

获取有标注的高质量医疗数据极其耗时且成本高昂。因此，本节旨在探究少量标注数据是否仍能带来良好的性能。具体而言，我们旨在分析哭声自监督学习适配在这种情况下能发挥怎样的作用，以及哪种监督微调方法能取得最佳性能。为此，我们使用随机抽取的神经损伤数据集标注数据子集，分别进行线性+批归一化和端到端微调实验。

在图2中，我们展示了使用两个模型进行神经损伤分类的结果：一个是在VGGSound上通过自监督学习预训练、但未经过SSL哭声适应阶段的模型（表3第3行），以及一个通过SSL哭声适应获得最佳性能的模型（表3最后一行）。

随着监督子集的规模减小，自监督学习（SSL）哭声适配模型（实线）的性能始终优于未适配模型（虚线），且优势幅度不断扩大。另一个值得注意的现象是，对于未适配模型，随着标注微调数据量的减少，端到端微调策略的性能下降得更快。这或许可以解释为什么在哭声诱因任务中，端到端适配的表现不如线性+批归一化（表4）：与神经损伤任务相比，该任务可用于微调的标注数据更少。

非常有趣的是，我们发现，即使只有 3%（几十个样本）的标注数据，使用哭声适配模型仍可达到约 70% 的曲线下面积。此外，仅使用 20% 的数据时，适配后的模型表现也显著优于我们的监督基线。这表明，哭声自监督学习适配具有巨大潜力，只需在监督微调阶段加入少量标注数据，就能获得令人满意的

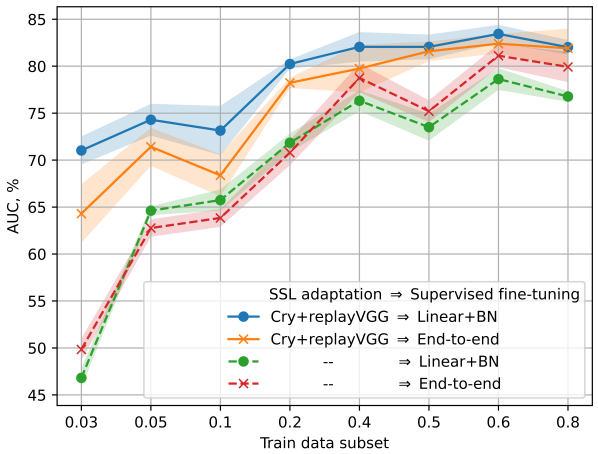


图2：在监督微调中使用部分带标签的神经损伤数据时的性能（Linear+BN 和端到端）。实线表示采用 SimCLR cry 适配的模型（表3最后一行），虚线表示未进行适配的相同模型（表3第3行）。阴影区域表示使用不同随机种子进行五次运行所得曲线下面积的标准误。

仅需在监督微调阶段引入少量标注数据，即可提升模型性能。

6. 结论

本文探讨了将大规模自监督预训练用于婴儿哭声分析，具体而言，是检测神经系统损伤和哭声诱因。我们发现，自监督学习预训练的效果显著优于传统监督预训练，而两者的效果都显著优于随机初始化训练。此外，在标注数据有限的情况下，我们发现，利用哭声专用未标注数据进行自监督学习适配，能显著减少监督微调阶段对标注数据的需求。我们还表明，通过自监督学习利用未标注哭声数据对编码器进行适配，可以显著提升神经系统损伤检测的下游性能。因此，我们认为，借助大量未标注哭声录音，这为研究开辟了一个充满希望的方向：仅使用目标人群中的少量标注哭声，就有可能训练出能够检测新疾病的分类器。

7. 致谢

我们谨向所有采集哭声数据的研究点负责人致谢：Uchenna Ekwochi 博士（尼日利亚埃努古州立大学教学医院）、Boma West 博士和 Datonye Briggs 博士（尼日利亚河流州立大学教学医院）、Peter Ubuane 博士（尼日利亚拉各斯州立大学教学医院）、Gabriel Variane 博士（巴西圣保罗圣卡萨慈悲医院）以及 Guilherme Sant'Anna 博士（加拿大蒙特利尔儿童医院）。我们还要感谢 Zhepei Wang 提供预训练模型并给予宝贵建议。最后，我们衷心感谢所有研究点的婴儿、他们的家人以及医疗和研究人员，感谢他们参与研究并协助采集哭声数据。

8. 参考文献

[1] O Wasz-H¨ockert 等人，《婴儿发声中特定含义的识别》，Experientia，第 20 卷，1964年。[2] Chunyan Ji 等人，《婴儿哭声分析与分类综述》，EURASIP《音频、语音与音乐处理》期刊，2021年。[3] Katarina Michelsson 等人，《足月窒息新生儿的疼痛哭声与晚期发现的相关性》，Acta Pædiatrica，第 66 卷，第 5 期，1977年。[4] Lichuan Liu 等人，《婴儿哭声语言分析与识别：一种实验方法》，IEEE/CAA 自动化学报，第 6 卷，第 3 期，2019年。[5] Gustavo Z Felipe 等人，《使用频谱图识别婴儿哭声的动机》，载于国际系统、信号与图像处理会议（IWSSIP）。IEEE，2019年。[6] Joanna J Parga 等人，《通过哭声声学分析定义和区分婴儿行为状态：肠绞痛是否会引起疼痛？》，《儿科研究》，第 87 卷，第 3 期，2020年。[7] Charles C Onu 等人，《用于围产期窒息基于哭声诊断的神经迁移学习》，INTERSPEECH，2019年。[8] Najla Al Futaisi 等人，《VCMNet：用于婴儿发声成熟度自动分析的弱监督学习》，载于多模态交互国际会议，2019年。[9] Turgut Ozseven，《利用不同的深度神经网络模型和手工设计特征进行婴儿哭声分类》，《生物医学信号处理与控制》，第 83 卷，2023。[10] Hemant A Patil 等人，《用于正常与病理性婴儿哭声分类的常数Q倒谱系数》，载于 ICASSP。IEEE，2022年。[11] Orion F Reyes-Galaviz 和 Carlos Alberto Reyes-Garcia，《一种利用神经网络处理婴儿哭声并识别新生儿病理状况的系统》，载于 SPECOM，2004年。[12] Qiuqiang Kong 等人，《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，2020年。[13] Andreas Triantafyllopoulos 和 Bj¨orn W Schuller，《任务与声学相似性在音频迁移学习中的作用：来自语音情感识别案例的见解》，载于 ICASSP。IEEE，2021年。[14] Edresson Casanova 等人，《面向 ComParE 2021 中 COVID-19识别任务的迁移学习与数据增强技术》，载于 INTERSPEECH，2021年。[15] Abdelrahman Mohamed 等人，《自监督语音表征学习：综述》，IEEE《信号处理专题期刊》，2022年。[16] Shu-wen Yang 等人，《SUPERB：语音处理通用性能基准》，INTERSPEECH，2021年。[17] Luyu Wang 等人，《迈向通用音频表征学习》，载于 ICASSP。IEEE，2022年。[18] Ting Chen 等人，《视觉表征对比学习的简单框架》，载于 ICML，2020年。[19] Ting Chen 等人，《大型自监督模型是强大的半监督学习器》，NeurIPS，第 33 卷，2020年。[20] Zhepei Wang 等人，《通过持续自监督学习新声音类别的表征》，IEEE《信号处理快报》，第 29 卷，2022年。[21] Janne Spijkervet 和 John Ashley Burgoyne，《音乐表征的对比学习》，ISMIR，2021年。[22] Matthew C McCallum 等，“面向音乐理解的音频表征监督学习与无监督学习”，*IS- MIR*，2022年。[23] Daniel S Park 等，“SpecAugment：一种用于自动语音识别的简单数据增强方法”，*INTER- SPEECH*，2019年。[24] CC Onu 等，“Ubenwa：基于哭声的出生窒息诊断”，*NIPS Workshop on Machine Learning for the Devel- opingWorld*，2017年。[25] Zhengyang Chen 等，“基于自监督学习的鲁棒说话人验证领域自适应”，载于*ICASSP*，2021年。[26] Suchin Gururangan 等，“预训练不止步：使语言模型适应领域和任务”，*ACL*，2020年。[27] Alexei Baevski 等，“wav2vec 2.0：语音表征自监督学习框架”，*NIPS*，第33卷，2020年。[28] Wei-Ning Hsu 等，“HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习”，*IEEE/ACM《音频、语音与语言处理汇刊》*，第29卷，2021年。[29] Sanyuan Chen 等，“WavLM：面向全栈语音处理的大规模自监督预训练”，*IEEE《信号处理专题期刊》*，第16卷，第6期，2022年。[30] Yuan Gong 等，“SSAST：自监督音频频谱图变换器”，载于*Proceedings of the AAAI Conference on Arti- ficial Intelligence*，2022年，第36卷。[31] Honglie Chen 等，“VGGSound：大规模视听数据集”，载于*ICASSP*，2020年。[32] Jort F. Gemmeke 等，“AudioSet：音频事件的本体与人工标注数据集”，载于*ICASSP*，2017年。[33] Jonathan Frankle 等，“训练批归一化且仅训练批归一化：卷积神经网络中随机特征的表达能力”，*ICLR*，2020年。[34] Moslem Yazdanpanah 等，“重新审视少样本迁移学习中批归一化的可学习仿射变换”，载于*IEEE/CVF Con- ferenceonComputerVisionandPatternRecognition*，2022年。[35] Diederik P Kingma 和 Jimmy Ba，“Adam：一种随机优化方法”，*ICLR*，2015年。[36] Harvey B Sarnat 和 Margaret S Sarnat，“胎儿窘迫后新生儿脑病：一项临床及脑电图研究”，《神经病学档案》，第33卷，10，1976年 no. 。[37] Bj¨orn Schuller 等，“INTERSPEECH 2016计算副语言学挑战：欺骗、真诚与母语”，载于*INTERSPEECH*，2016年。[38] Florian Eyben 等，“OpenSmile：慕尼黑通用快速开源音频特征提取器”，载于*Proc. of the 18th ACM international conference on Multimedia*，2010年。[39] Robert M. French，“联结主义网络中的灾难性遗忘”，《认知科学趋势》，第3卷，第4期，1999年。[40] Anthony Robins，“灾难性遗忘、复习与伪复习”，《联结科学》，第7卷，第2期，1995年。