使用无监督声音分离改进鸟类分类

汤姆·丁顿，斯科特·威斯дом，约翰·R·赫希

谷歌研究

摘要

本文探讨了鸟类歌声录音中的物种分类问题。大量可用的鸟类野外录音为利用机器学习自动追踪鸟类种群提供了机会。然而，这也带来了一个问题：这些野外录音通常包含显著的环境噪声和重叠的鸣叫声，干扰了分类。用于物种识别的广泛训练数据集通常也未能标记背景物种。这导致分类器忽略了信噪比低的鸣叫声。然而，无监督声音分离的最新进展，如混合不变训练（MixIT），能够从这些嘈杂录音中学习高质量的鸟类歌声分离。在本文中，我们展示了专门针对鸟类歌声数据训练的 MixIT 模型在分离质量上的提升，其重建混合物的 SI-SNR 改善超过 5 dB，优于通用音频分离模型。我们还展示了在三个独立数据集上，下游多物种鸟类分类器的精度提升。 最佳分类器性能是通过在分离的通道和原始音频上获取最大模型激活值来实现的。最后，我们记录了额外的分类器改进，包括分类学分类、随机低通滤波器增强和额外的通道归一化。

索引词—源分离、无监督学习、分类算法、生态学

1. 引言

生物声学机器学习是一个不断发展的领域，有潜力改变我们对自然环境的理解。最近的成功包括发布像 BirdNET 和 Merlin Sound ID 这样的应用程序，允许用户识别他们周围环境中的鸟类。与此同时，像雨林连接这样的组织使用音频数据来识别濒危物种的关键栖息地，并防止非法砍伐和偷猎。

声学数据比相机陷阱提供了更多观察动物的机会，因为相机陷阱存在视野狭窄和探测范围有限的问题。鸟类、鲸鱼以及无数其他动物使用发声作为安全有效的交流方式，并标记领地。这些发声动物还充当指示物种，使我们能够进一步推断栖息地和生态系统健康状况，包括植物和昆虫食物来源、捕食者以及整体生物多样性。几十年来，像北美繁殖鸟类调查[1]这样的工作一直依靠专家听众追踪鸟类种群，并为栖息地保护和恢复决策提供信息。廉价音频录音工具的可用性大大提高了采样覆盖率，但也产生了需要引入自动化处理的大量数据。

然而，鸟类物种分类器在处理原始声音景观数据时仍然存在困难，尤其是在背景噪音显著的情况下。在噪音中或清晨合唱时，当发声活动达到顶峰且许多物种同时鸣叫时，这种困难会更加明显。这种困难的原因在于，用于鸟类歌声的训练数据通常在音频片段中只标记一个主要物种的标签，即使有多个物种在鸣叫，也会将较轻的背景鸣叫声未标记。当将分类器应用于声音景观时，这会产生一个领域偏移，因为我们想要识别录音中的所有物种，而不仅仅是最突出的物种。例如，像红翅黑鸟（rewbla）这样非常常见的物种，可能会频繁出现在其他物种录音的背景中，导致分类器学习到“安静”的红翅黑鸟鸣叫声通常是未标记的。

近来，在利用含噪训练数据进行无监督音频源分离方面取得了显著进展。混合不变训练（MixIT）[2] 创建了一个分离模型，能够从单声道录音中分离出各个声音。关键在于，训练 MixIT 模型不需要干净音频源，这与大多数之前的系统不同。这为生物声学提供了一个明确的机会：借助 MixIT 分离模型，我们可以在混乱的声景中分离出单个发声，抑制过度的背景噪声，从而提高分类性能。

贡献：在本文中，我们展示了当将无监督分离模型与鸟鸣分类器结合时，分类性能的改进。我们在三个不同的评估数据集上，在多种分类指标上展示了改进效果。

我们也展示了相对于先前发表的工作，在基础分类模型上的若干改进。最值得注意的是，我们引入了分类训练，其中我们为物种分类的每个级别训练分类头：物种、属、科和目，以及一个二进制鸟类检测头。这使模型能够在区分密切相关物种之间有时细微的差异之前学习更高级别的标签。在层次分类[3]的背景下，我们训练了一个全局分类器，尽管我们仅使用更高级别的分类类别来改进物种问题的学习。我们还记录了来自额外的梅尔谱图清理、低通增强和集成多个分类器的收益。

2. 方法

2.1. 训练数据活动检测

所有训练数据都是弱标签的，为录音中的主要物种提供一个全局标签，尽管在可能存在长时间未出现该标签物种的时期。我们使用活动检测器对训练数据集进行预处理，以选择可能包含该标签物种的短窗口。为了选择一个窗口，我们为每条录音创建一个对数梅尔谱图，并计算每个帧的能量。然后我们应用小波峰值检测器[4]（scipy 的 find peaks cwt）来找到高能量的帧。我们围绕每个峰值帧提取一个 6 秒的窗口，并选择最多五个

1 为简洁起见，我们用 6 个字母的 eBird 代码来指代单个物种。

按峰值帧能量降序排列的窗口。这个过程可以可靠地找到标记物种，但偶尔会选中背景发声或人声。

2.2. 分离模型概述

最近，提出了混合不变训练（MixIT）[2] 作为一种在缺乏真实参考源的自然声混合物上训练声音分离模型的方法。基本思想是从两个训练参考混合物中创建混合物的混合物（MoMs），从 MoM 中估计 M 个源，将每个源分配给两个参考混合物中的一个，并计算每个参考混合物与其分配源之和之间的重建损失。源到混合物的分配选择以最小化此重建损失。然后通过最小化给定源分配的重建损失来优化模型。实际上，模型可以通过为某些输出生成接近零的信号来产生少于 M 个输出源。

更详细地说，两个参考混合物 x∈ Rand x∈ R 相加创建一个 MoM，¯x = x+ x。分离模型 fpredicts M 个源 ˆs ∈ Rfrom the MoM: ˆs = f(¯x)。使用参考混合物 x 和分离源 ˆs 之间的判别信号级损失 L，MixIT 损失 [2] 估计一个混合矩阵 A ∈ B：

$$L(\{x\}, \hat{s}) = \min_{A \in B^{2\times M}} \sum_{n=1}^{2} L(x, [A\hat{s}]) \quad (1)$$

其中 B 是所有 2 × M 二进制矩阵的集合，其中每一列的和为 1。该矩阵将每个分离的源ˆs 分配给一个参考混合 x。对于信号级损失 L，我们使用负阈值信噪比损失[2, 5]：

$$L(y, \hat{y}) = -10 \log \frac{||y||}{||y - \hat{y}|| + \tau ||y||} \quad (2)$$

其中 τ = 10 作为软阈值，将损失限制在信噪比（SNR）范围内。这个阈值防止已经很好地分离的样本在训练批次中主导梯度。我们发现 SNR= 30 dB 是一个较好的值。

分离模型是一种基于掩码的架构。首先，输入音频通过可学习的基进行变换[6]。这些系数被输入到改进的时域卷积网络（TDCN++）[7]中，该网络通过 sigmoid 激活函数预测 M 个掩码。这些掩码与分析系数相乘，并使用可学习的合成基通过重叠相加方法合成音频波形。最后，应用混合一致性投影[8]，该投影约束源信号相加等于原始输入。

2.3. 数据增强

数据增强众所周知是优秀鸟类声音分类器的一个基本组成部分[9]。在分类器训练过程中，我们对训练数据应用以下增强：

随机时间偏移：我们从训练集中的每个 6 秒示例中选择一个随机的 5 秒训练窗口。

随机增益：对于每个训练示例，我们进行峰值归一化到 0.05 到 0.75 之间的均匀随机值。

示例混合：以 50%的概率，我们混合一个带有随机增益的第二标记训练示例。目标标签集是两个示例标签的并集。

噪声混合：我们使用来自以下来源的噪声录音：(1) 2018 DCASE Bird Audio 的负类检测挑战[10]，(2)从 BBC 自然音效库中选择类别，以及(3)从 Common Voice 数据集[11]中随机选择录音集。在 75%的训练样本中混合噪声，信噪比(SNR)在 0 到 40 dB 之间均匀随机选择。有 10%的概率，我们抑制标记音频和标签，仅用噪声进行训练。

随机低通滤波：除了降低增益，远离音频源也会产生低通效果。我们通过缩放梅尔谱图中的频段来模拟随机低通滤波器的应用。

2.4. 分类器架构

对于分类器前端，我们计算增强音频的梅尔谱图，并应用 PCEN [12]，该技术已被广泛观察到有助于分类 [13]。我们还对 PCEN 梅尔谱图应用了最终的通道归一化。对于这次最终归一化，我们计算通道均值和方差，并去除所有超过均值一个正偏差的离群值。然后我们重新计算排除离群值后的均值和方差，并对通道进行归一化。

分类器模型使用直接应用于 PCEN 梅尔谱图的 EfficientNet-B0 主干网络 [14]。在 EfficientNet 之后，我们应用 AutoPool [15]来减少时间维度，并投影到 1280 维的隐藏空间。对于物种分类学的每个级别（物种、属、科和目），我们使用单独的分类头，标签自动从物种中派生。属、科和目的输出损失权重为 0.1。这些输出仅在训练期间使用。

所有输出头均使用二元交叉熵损失（允许每个标签独立激活），标签平滑因子为 0.1。每个模型从零开始训练五次，通过在 logit 域中对每个物种的输出概率进行平均来作为集成使用。

2.5. 分类分离音频

为了结合分离和分类模型，我们将分离模型应用于输入音频窗口以获得 M 个输出通道。然后我们将分类模型应用于每个分离通道和原始音频，并取每个物种的最大概率。我们发现将原始混合物作为通道包含在内显著提高了结果，如表 2 所示。

3. 实验

3.1. 鸟鸣训练数据集和处理

为了训练分离模型和分类器，我们使用来自 Xeno-Canto [16]和 Macaulay Library [17]的数据组合。这些源文件长度从几秒到超过五分钟不等，且为弱标签，整个文件只有一个标签，但没有分割信息。一些 Xeno-Canto 文件还包括背景标签，指示录音中出现的主要标签物种以外的其他物种。在这种情况下，仍然没有分割信息可用，对于记录者可能未知或难以识别的鸟类可能仍然未标记。

每个训练数据集专门针对特定的目标物种。

我们为每个物种选择最多 250 个训练录音，优先选择 Macaulay 录音和用户评分高且带有背景标签的 Xeno-Canto 录音。我们额外包含 250 个文件

从目标物种集之外的物种中随机选择。在分类器训练期间，这些附加文件中的片段不提供物种标签，尽管任何相关的分类学标签（属、科、目）仍然包括在内，用于计算分类学损失。

3.2. 评估数据集

我们考察了三个独立的评估数据集，使用两个不同的物种集，覆盖纽约州上州和加利福尼亚州的内华达山脉。

Sapsucker Woods 数据集（SSW）是 2019 年 BirdClef 竞赛[9]的测试集，包含在纽约州伊萨卡录制的 40k 个 5 秒片段。该数据集包含 70 个物种，所有物种在数据集中至少出现五次。音频由 SWIFT 自动录音单元（ARUs）录制，并由康奈尔鸟类学实验室的专家手动标注。我们包含 SSW 数据集，以便与现有方法进行一些比较。

High Sierras 数据集（HSN）在内华达山脉的高海拔地区收集，并在 BirdClef 2021 竞赛[18]中描述。在从标注录音中删除静音区域后，该数据集包含 4928 个标注的 5 秒片段，有 18 个物种出现，其中 16 个物种至少有 5 次鸣叫。标注录音是在早上 4 点到 8 点之间进行的，涵盖了晨间合唱。

卡普尔斯数据集（CAP）来自加利福尼亚州塔霍湖附近的卡普尔斯流域。2017 年和 2018 年，使用 ARU 收集了来自 80 个点的音频，作为一项研究受控燃烧对生物多样性的影响的项目的一部分。为了创建评估集，加利福尼亚科学学院的志愿者选择了 3 分钟的片段进行标注，提供了时间分割和标签的置信度。出于评估目的，丢弃了低置信度标签和静音区域。最终数据集包含 2944 个标注的 5 秒片段，共 42 个物种。其中 36 个物种至少有 5 个发声。

3.3. 评估指标

我们使用一系列指标来理解模型质量。

类别平均平均精度（CMAP）和标签加权标签排序平均精度（lwlrap）是密切相关的指标，适用于多类别多标签场景，它们将平均倒数排名统计推广到多标签场景。CMAP 是许多 BirdCLEF 竞赛的目标指标[9]，而 lwlrap 是 DCASE 2019 音频挑战的目标指标[19]。简而言之，CMAP 是每类精度分数的平均值，而 lwlrap 是每个样本精度分数的平均值。对于观察值较少的物种，每类精度非常嘈杂，因此我们仅使用数据集中至少有五个观察值的物种来计算 CMAP。请注意，lwlrap 反映了数据集中的类别不平衡。

dsensitivity 指数是对 AUC 的重新缩放（将每个标签视为一个具有公共阈值的独立二元分类问题）。定义 d=

$\sqrt{2}Z(AUC)$，其中 Z 是标准高斯分布的 CDF。它衡量正负标签的整体分离质量。

最后，top-1 精度不是多标签指标，但仍然有助于评估模型质量。

3.4. 模型训练细节

分离模型在 704 个物种的数据集上进行训练

包括纽约和内华达山脉物种以及 BirdCLEF 2019 挑战中包含的南美其他物种，音频片段采样率为 22.05 kHz，时长为 6 秒。学习到的滤波器组使用 1 毫秒的窗口大小、0.5 毫秒的步长和 256 个系数，模型产生 M = 4 个分离的输出通道。

为内华达山脉（基于 89 个物种的数据集）和纽约（基于 94 个物种的数据集）分别构建了两个独立的分类器集合。输入梅尔频谱的帧率为 100 Hz，帧长为 0.08 秒，频率范围为 60 至 10,000 Hz，输出通道数为 160。单个分类器使用 Adam 优化器进行 200k 步训练，学习率为 0.01，批处理大小为 64。每个模型使用 32 个 v2 TPU 大约需要 3.5 小时进行训练。更长的训练时间会导致评估的 CMAP 指标下降。对于 Xeno-Canto 的训练示例，我们将背景标签中任何物种的损失设置为零，以避免惩罚正确的识别结果。

3.5. 分离结果

为了展示在匹配的音频数据域上训练分离模型的有效性，我们将第 2.2 节中描述的、在 Xeno-Canto 训练数据上训练的分离模型与一个类似的分离模型进行比较。该类似模型使用 MixIT 在 AudioSet [20] 的 5800 小时原始音频数据上训练。评估数据由 3856 个 MoMs 组成，这些 MoMs 是从保留的 Xeno-Canto 评估数据的 6 秒片段中创建的。为了衡量性能，我们使用 MoMi 分数 [2]，它是估计混合物上的平均尺度不变信噪比 [21] 改善（SI-SNRi）。对于每个示例，通过找到每个分离源到两个参考混合物之一的最佳分配，并求和这些源来计算两个估计混合物。

目前可用的用于比较的最接近的预训练 AudioSet 模型具有 M = 4 个输出，是在 16 kHz 音频上训练的，并使用可学习的基窗口长度为 2.5 ms。为了进行公平的比较，我们也使用 2.5 ms 窗口长度而不是 1 ms 训练了 Xeno-Canto 模型，并通过将它们降采样到 16 kHz 采样率（SR）并使用 16 kHz 参考进行评分来评估 22.05 kHz 模型输出。表 1 显示了与预训练 AudioSet 模型相比，Xeno-Canto 训练模型的性能。这些结果表明，针对分离训练匹配数据而不是 AudioSet 的通用音频数据，确实为该应用提供了更好的分离模型，并在 16 kHz 评估数据上 MoMi 指标提升了超过 5 dB。将窗口大小从 2.5 ms 减小到 1.0 ms 进一步提升了近 1 dB 的性能。

表 1：Xeno-Canto MoM 评估集上的分离结果。

训练数据	输入采样率	评估采样率	窗口长度	MoMi
AudioSet	16 kHz	16 kHz	2.5 ms	4.4 dB
Xeno-Canto	22.05 kHz	16 kHz	2.5 ms	9.6 dB
Xeno-Canto	22.05 kHz	16 kHz	1.0 ms	10.4 dB
Xeno-Canto	22.05 kHz	22.05 kHz	1.0 ms	10.5 dB

分离而言，在 Xeno-Canto 上训练确实提供了更好的分离模型，并在 16 kHz 评估数据上 MoMi 指标提升了超过 5 dB。将窗口大小从 2.5 ms 减小到 1.0 ms 进一步提升了近 1 dB。

3.6. 与先前分类结果的比较

在 BirdCLEF 2019 [9] 竞赛 中，最佳模型在 SSW 数据集上实现了 0.231 的 CMAP，当使用来自 SSW 验证集的强标记数据进行训练时，这一数值上升到了 0.407。我们的分类模型没有在验证集上进行训练，但拥有稍大的训练集，每个物种最多有 250 条录音，而 BirdCLEF 挑战中每个物种只有 100 条录音。纽约分类器单独使用时平均 CMAP 为 0.242，作为五个模型的集成时 CMAP 为 0.304。

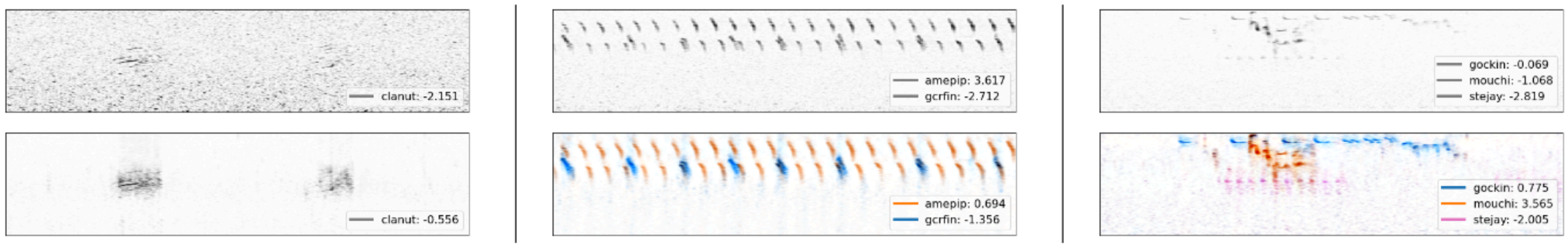

图 1：分离+分类示例。上方的图显示了原始音频的 PCEN 梅尔谱图，下方的图显示了分离音频的 PCEN 梅尔谱图，其中分离的通道使用颜色编码显示。图例给出了真实物种的集成逻辑值。左侧：High Sierras 数据集中的 Clarke’ s Nuthatch，说明在分离通道中简单的噪声抑制通常可以提高低信噪比的孤立鸣叫的逻辑值。中间：High Sierras 数据集中一个具有挑战性的双源示例。

正确：Caples 数据集的三种物种分离。

表 2：比较 Separate+Classify 方法。 “Mix Only”分数是原始音频的集成分类器分数。 “Separation”分数是 4 个分离音频通道中每个物种的最大概率。 “Mix+Separation”取分离通道和原始音频中的最大物种概率。分类由 EfficientNet-B0 模型的集成执行。

	Sapsucker Woods			
CMAP lwlrap dTop-1 Mix Only 0.304 0.431 1.117				
0.398 Separation 0.268 0.413 1.116 0.360				
Mix+Separation 0.306 0.441 1.123 0.397				
	Caples			
仅混合 0.334 0.569 1.144 0.496 分离 0.327 0.581				
1.154 0.506 混合+分离 0.341 0.590 1.155 0.517				
	High Sierras			
仅混合 0.527 0.531 1.149 0.432 分离 0.548 0.548				
1.149 0.448 混合+分离 0.560 0.560 1.153 0.451				

3.7. 结果讨论

如表 2 所示，评估结果表明，通过结合分离和分类模型，在多个数据集上未校准的分类指标几乎都有均匀的提高。此外，将原始的噪声混合作为附加通道处理的效果比单独使用分离后的音频更好。表 3 展示了对分类器的消融研究结果，展示了各种组件和架构的相对贡献。

图 1 展示了三个示例，更多示例可在网上找到。通过检查具体示例，我们观察到分离带来的几项优势。首先，在许多情况下，经过背景噪声去除后，孤立（不重叠）的低信噪比发声的概率会显著提高（图 1 左侧）。弱标记的训练数据包含许多未标记的低信噪比发声，因此模型被训练为忽略这些发声。分离后，PCEN 处理使这些发声清晰地出现在前景中。其次，常见物种与其他物种的混合物往往会降低常见物种的概率，而分离处理后概率会提高。这可能是因为常见物种经常出现在其他物种的训练录音中，导致其概率较低。第三，我们最初的动机是分离具有重叠发声的复杂场景，例如晨间合唱。我们发现模型能够成功分离重叠发声，即使在许多情况下鸣禽的频率范围重叠（图 1 中部、右侧）。

表 3：分类器消融。对于每个分类器，我们报告了五个独立模型的平均指标。我们还提供了 EfficientNet B0 集成分数以供比较。

	Caples		High Sierras	
CMAP lwlrap Top-1 CMAP lwlrap Top-1 B0 (Ensemble) 0.334 0.569				
0.496 0.527 0.531 0.432 B0 (Mean) 0.284 0.517 0.458 0.479 0.483				
0.392 -Taxo Loss 0.272 0.491 0.434 0.465 0.469 0.376 -Lowpass 0.283				
0.467 0.423 0.401 0.398 0.319 -Channel Norm 0.270 0.484 0.437				
0.438 0.447 0.362 B0 (Mean) 0.284 0.517 0.458 0.479 0.483 0.392 B1				
0.275 0.503 0.448 0.477 0.479 0.396 B2 0.268 0.490 0.437 0.486 0.477				
0.390				

使用分离也产生了一些缺点。首先，我们经常观察到，在给定录音中，最突出的物种的概率在分离后会降低。这可能是由于丢失了额外的音频上下文。将原始混合音与分离的声道一起使用可以缓解这种效果。其次，如果我们只在分离的声道上应用分类，评估分数会下降。这可能是由于过度分离、领域偏移或音频上下文丢失。第三，过度分离偶尔会发生，有时会从一首较长的歌曲中分离出几个音符。孤立的音符可能与另一种物种的叫声相似，导致错误分类。

4. 结论

我们通过将下游分类器与特定领域的 MixIT 分离器相结合，证明了在生物声学中训练该分离器的实用性。这项工作可以清晰地扩展到其他生物声学问题，例如蛙鸣合唱和海洋声音景观分析。我们希望找到更好的方法来结合分离和分类系统。例如，将分离应用于分类器训练数据可能会提高分类器的质量。将源分离到特定通道中，也将允许通过将鸟类、昆虫和两栖动物发声分离到不同通道中来创建更好的声学指标。

5. 致谢

我们感谢 Mary Clapp、Jack Dumbacher、Durrell Kapan 和加利福尼亚科学学院提供的 Caples 和高山地区数据集，以及他们对分类器质量的广泛定性反馈。我们也要感谢 Stefan Kahl、Holger Klinck 和康奈尔鸟类学实验室的保护区生物声学中心提供的 Sapsucker Woods 数据集。所有模型均使用来自 Xeno-Canto 和麦克劳利图书馆的数据进行训练；没有它们，这项工作将不可能完成。

2https://bird-mixit.github.io

6. 参考文献

[1] Marie-Anne R Hudson, Charles M Francis, Kate J Campbell, Constance M Downes, Adam C Smith, 和 Keith L Pardieck, “北美洲繁殖鸟类调查的作用”
《观察》，《信天翁：鸟类学应用》，第 119 卷，第 3 期，第 526-545 页，2017 年。

[2] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J Weiss, Kevin Wilson 和 John R Hershey, “使用混合不变训练的无监督声音分离,” 载于《神经信息处理系统进展》, 2020.

[3] 卡洛斯·N·希拉和亚历克斯·A·弗雷塔斯， “跨不同应用领域的层次分类综述，” 数据挖掘与知识发现，第 22 卷，第 1 期，第 31-72 页，2011 年。

[4] Pan Du, Warren A Kibbe, and Simon M Lin, “通过结合基于小波变换的模式匹配改进质谱峰值检测,” bioinformatics, vol. 22, no. 17, pp. 2059–2065, 2006.

[5] Zhong-Qiu Wang, Hakan Erdogan, Scott Wisdom, Kevin Wilson, and John R Hershey, “用于语音分离和增强的顺序多帧神经波束成形,” in
IEEE Spoken Language Technology Workshop (SLT) 会议论文集，2021 年。

[6] Yi Luo and Nima Mesgarani, “Conv-TasNet: 超越理想时频幅度掩蔽的语音分离”
“tion，” IEEE/ACM 语音、语言和音频处理汇刊，第 27 卷，第 8 期，第 1256-1266 页，2019 年。

[7] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, 和 John R Hershey, “Uni-
"通用声音分离”，在 IEEE 声学与音频信号处理应用研讨会（WASPAA）2019 年会议论文集中。

[8] Scott Wisdom, John R. Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A. Saurous, “可微一致性约束用于改进深度学习”
语音增强，” 在 IEEE 国际声学、语音与信号处理会议（ICASSP）2019 年会议论文集中，第 900-904 页。

[9] 斯特凡·卡尔，法比安-罗伯特·斯托尔，埃尔韦·戈埃，埃尔韦·格洛廷，罗伯特·普兰克，威廉-皮尔·维林加，和亚历克西斯·乔利， “鸟 CLEF 2019 概述：大规模鸟类识别在”
"声音景观"，载于 CLEF 2019 会议与评估论坛实验室会议论文集。CEUR，2019 年，第 2380 号，第 1-9 页。

[10] 弗朗茨·贝格尔，威廉·弗里林格，保罗·普里姆斯，和沃尔夫冈 Reisinger, “鸟类音频检测 - DCASE 2018,” 技术报告，DCASE2018 挑战赛，2018 年 9 月。

[11] 罗萨娜·阿尔迪拉，梅根·布兰森，凯莉·戴维斯，迈克尔·亨-Pretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, 和 Gregor Weber, “Common voice: 一个多语言语音语料库,” 在 LREC, 2020.

[12] 王宇轩，Pascal Getreuer，Thad Hughes，Richard F Lyon，and Rif A Saurous， “可训练的前端用于鲁棒和远场关键词识别”，在 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集中。IEEE，2017，第 5670-5674 页。

[13] Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee、Andrew Farnsworth、Steve Kelling 和 Juan Pablo Bello， “每通道能量归一化：原因和方法，” IEEE 信号处理信札，第 26 卷，第 1 期，第 39-43 页，2018 年。

[14] 明星·谭和 Quoc Le， “EfficientNet: 重新思考模型"卷积神经网络的缩放，” 在机器学习国际会议。PMLR, 2019, pp. 6105– 6114.

[15] Brian McFee、Justin Salamon 和 Juan Pablo Bello， “Adap-用于弱标签声音事件检测的激活池化算子，” IEEE/ACM 音频、语音和语言处理汇刊，第 26 卷，第 11 期，第 2180-2193 页，2018 年。

[16] “Xeno-Canto：分享来自世界各地的鸟类声音，” https://www.xeno-canto.org/, 访问时间：2021-10-05。

[17] “康奈尔鸟类学实验室的麦克劳利图书馆，” https://www.macaulaylibrary.org/, 访问时间：2021-10-05。

[18] S Kahl, T Denton, H Klinck, H Glotin, H Go¨eau, WP Vellinga, R Planqu´e 和 A Joly， “鸟鸣识别综述：声音景观录音中的鸟鸣识别，” CLEF 工作笔记，2021。

[19] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis 和 Xavier Serra， “使用噪声标签和最少监督进行音频标记”，在 DCASE2019 工作坊，美国纽约，2019 年。

[20] Jort F Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, 和 Marvin Ritter, “Audio Set: 一个音频事件的本体和人类标注数据集,” 在 IEEE 国际声学、语音与信号处理会议 (ICASSP) 会议论文集中, 2017, pp. 776–780.

[21] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, “SDR–半成品还是成品?,” 在 IEEE 国际声学、语音与信号处理会议 (ICASSP) 2019 年会议论文集，第 626–630 页。