arXiv:2305.05591v1 [cs.SD] 9 May 2023

AUDIOSLOTS：一种以槽位为中心的音频分离生成模型

Pradyumna Reddy[*]　　*Scott Wisdom*、*Klaus Greff*、*John R. Hershey*、*Thomas Kipf*

伦敦大学学院　　Google Research

摘要

近期的一系列研究表明，以对象为中心的架构适用于视觉领域中的无监督场景分解。受这些方法启发，我们提出AudioSlots，这是一种用于音频领域盲源分离的以槽位为中心的生成模型。AudioSlots 采用置换等变的编码器和解码器网络构建。基于 Transformer架构的编码器网络学习将混合音频频谱图映射为一组无序的独立源嵌入。空间广播解码器网络则学习根据源嵌入生成源频谱图。我们使用置换不变损失函数，以端到端方式训练该模型。在 Libri2Mix 语音分离任务上的结果表明，这一方法颇具潜力，可作为概念验证。我们详细讨论了该方法的结果和局限性，并进一步概述了克服这些局限性的潜在途径以及未来的研究方向。

索引词— 语音分离，以对象为中心的表示

1. 引言

近年来，基于神经网络、面向集合结构化数据的架构，以及学习将非结构化输入映射到集合结构化输出空间的架构，都受到了广泛研究。尤其是在视觉领域，以槽位为中心的或以对象为中心的架构支撑了目标检测 [1]和无监督目标发现 [2, 3]领域的最新进展。

这些以对象为中心的架构内置了置换等变性的归纳偏置，因此非常适用于音频分离任务。本文将这些架构的核心思想应用于声音分离问题：在无法获取有关声源或混合过程的特权知识的情况下，从混合音频信号中分离出音频源。声音分离本质上是一个基于集合的问题，因为声源的排列顺序是任意的。

我们将声音分离表述为一个置换不变的条件生成建模问题：学习将混合音频频谱图映射到一组无序的、相互独立的源频谱图。我们的方法 AudioSlots 将音频分解为每个声源各自的潜变量，随后将其解码为各自的源频谱图。该方法基于 Transformer 架构 [4]构建，采用具有置换等变性的编码器和解码器函数，因此不受声源潜变量（“槽位”）排列顺序的影响。

为评估这种架构的潜力，我们使用基于匹配的损失训练 AudioSlots，使其能够从混合音频信号中生成分离后的声源。我们在 Libri2Mix [5] 的一个简单双说话人语音分离任务上展示了该方法。

[*]在 Google 实习期间完成。

虽然我们的结果主要是对这一想法的概念验证，但我们发现，基于槽位的生成模型在声音分离方面展现出潜力，同时也面临一些挑战：当前版本的模型难以生成高频细节，需要依赖启发式方法拼接独立预测的音频片段，而且训练时仍需要真实参考音频源。我们相信这些挑战能够在未来的工作中得到解决，并在本文中概述了可能的研究方向。

2. 相关工作

我们的工作探索了一种用于声音分离的新型基于集合的生成建模方法。下文将简要概述近期基于学习的声音分离方法，以及其他领域中采用的基于集合（或以槽位为中心）的神经网络架构。

声音分离。针对监督式声音分离，研究者提出了多种神经网络方法，这些方法在声音重建机制和整体架构方面各不相同。**基于掩码的方法** 通过预测分离掩码来重建声音，并将其应用于输入音频信号的分析／合成基表示（例如短时傅里叶变换或学习得到的基），如 [6, 7, 8, 9, 10, 11, 12, 13]。另一种**直接重建**方法则估计源信号或其频谱，而不显式估计掩码 [14, 15]。

研究者已提出多种用于声音分离的通用架构，包括循环网络 [16], 卷积网络[9], U-Net [14], 注意力网络 [17], 及其组合。这些方法通过在训练过程中使用置换不变损失 [7, 8]来处理输出声源顺序任意的问题。

一些方法更进一步，在架构层面解决排列不变性问题，生成与每个声源对应的无序表示。深度聚类和深度吸引子网络 [6, 18, 7], 采用排列等变架构，以类似注意力的方式处理每个时频单元的嵌入。

我们的方法使用基于槽位的注意力机制为每个声源生成一个嵌入，并采用类似 NeRF 的 [19] 方法对其进行解码，直接估计声源频谱图。这与基于掩码的方法不同，采用的是直接预测；而利用 NeRF 架构进行直接预测的方法也颇具新意。基于槽位的注意力机制有别于以往用于声音分离的注意力网络，与深度聚类的关系则更为密切。不过，我们的注意力方法处理的是更高层次的频谱图区域，而非单独的时频单元，并且使用通用注意力机制，而不是简单的基于亲和度的方法。

近期的无监督方法，例如混合不变训练（MixIT） [20], 仅使用混合音频进行训练。虽然我们的工作仅探讨使用真实隔离参考声源进行监督式声音分离，但像 MixIT 这样的训练设置

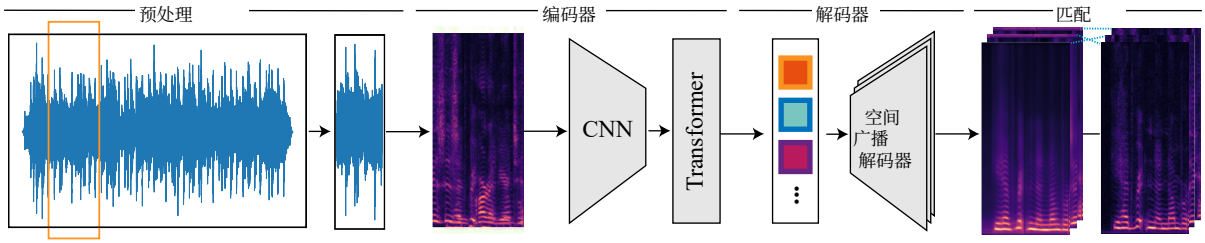

图 1：**架构概览**。首先对输入波形进行裁剪，并将其转换为频谱图。随后，神经网络将频谱图编码为一组置换不变源嵌入 $s_{1...n}$，再将这些嵌入解码，生成一组独立的源频谱图。整个处理流程使用真实源频谱图进行监督，并采用基于匹配的置换不变损失函数。

与我们的方法相互独立，值得在未来的工作中进一步探索。

以槽位为中心的神经网络。长期以来，研究人员一直在研究处理无序特征集的神经网络 [21, 22, 23] 。与我们的工作最相关的是根据输入数据生成无序输出集的方法 [24,25, 1], 以及使用一组潜变量（“槽位”）来建模输入中置换不变方面的方法 [2, 3], , 例如视觉场景中的物体。我们将后一类称为以槽位为中心的神经网络。在视觉领域，这类模型构成了许多现代场景理解方法的基础，包括目标检测 [1], 、全景分割 [26], 和无监督目标发现 [2, 3, 27]。在我们的工作中，我们展示了以槽位为中心的生成模型同样有望用于音频处理中的组合式置换不变任务，具体而言，就是从混合音频中分离出独立的音频源。

3. 方法

我们提出了一种用于监督式声音分离的广义置换不变训练框架。与以往方法不同，我们从生成式视角处理声源分离任务。我们方法的主要目标是将输入音频投影到一组嵌入向量中，每个向量代表输入中的一个不同声源。随后，以置换不变的方式利用这些嵌入向量生成各个声源的幅度频谱图。整个处理流程使用真实源频谱图进行监督，并采用置换不变损失。在本节其余部分，我们将详细介绍训练处理流程中的不同步骤。

预处理：在训练期间，给定一段混合音频波形，我们首先随机截取一个 0.5 秒的音频片段。随后，按照 [28], 我们使用窗长为 512、 hop 长为 125 的短时傅里叶变换，将片段转换为频谱图。然后，将复数频谱图的绝对值以 0.3 次幂进行非线性缩放，以突出较小的数值。缩放后的绝对值将作为下一步的输入。

编码：为了推断声源嵌入，首先使用 ResNet-34 网络 [29] 将输入频谱图编码为 32×8 网格的编码“图像”特征 z。我们在 ResNet 根块中采用较小的步幅，以保留更高的空间分辨率。接着，使用一个包含 4 层的 Transformer 将 z 映射为声源嵌入 $s_{1...n}$，其中 n 为声源数量，且 $s_i \in R^d$。与原始表述 [4],不同，在每个 Transformer 层中，我们先对查询向量 q 执行自注意力操作，然后对自注意力步骤的输出与 z执行交叉注意力。我们在自注意力和交叉注意力步骤中都使用相同的变量作为键和值。维度为 4096 的初始查询 q 通过反向传播学习，与 DETR [1] 类似。

解码：我们使用空间广播解码器 [30]，从 $s_{1...n}$生成各个源频谱图。首先，将每个嵌入s_i 复制到整个二维网格上，创建形状为 $F \times T \times d$ 的张量 g_i，其中 F，T分别表示输出频谱图的频率 bin 和时间步。然后，将带有傅里叶特征的 positional embeddings [31]附加到 g_i 上，使其形状变为 $F \times T \times (d + e)$，其中 e 是每个傅里叶特征嵌入的维度。随后，对 g_i 的 $F \times T$ 中的所有向量应用一个共享参数的全连接网络，得到一组频谱图，每个频谱图的形状为 。请注意，该解码器与 NeRF 模型类似：NeRF [19], 同样以位置编码为输入，并通过学习一个全连接网络来生成依赖于坐标的输出。在我们的模型中，我们直接生成频谱值，并将其排列在二维图像网格上作为输出；此外，还使用相应的潜在源嵌入 s_i 对网络进行条件化，以生成每个频谱图。

目标： 由于源分离问题具有排列不变性，预测的频谱图不会按照预先确定的顺序生成。因此，我们需要将估计的频谱图与真实频谱图匹配，以计算网络的损失。在所有真实频谱图与估计频谱图之间可能的匹配方式中，我们希望找到重建误差最小的最优分配。我们建议使用匈牙利匹配算法来解决这一分配问题，因为它速度快、准确度高，且能够高效处理大量项目（尽管我们的实验只使用声源数量较少的数据集）。该算法会为每一对可能的项目分配代价或权重，然后选出使相关总代价或总权重最小的配对。最后，将匹配后的真实频谱图与估计频谱图之间的均方误差最小化，作为优化网络参数的训练目标。

源分离： 测试时，给定一个输入混合波形，我们首先将其拆分为多个互不重叠、长度为 0.5 秒的波形。如果输入波形的长度不能被 0.5 秒整除，我们会在末尾添加零填充。这些波形按照上述方法进行预处理，然后作为输入传递给网络。该网络使用上述处理流程进行训练，以估计各个源的频谱图的绝对值。首先，将估计的频谱图进行 1/0.3 次幂运算，以还原预处理时进行的缩放。随后，使用这些重新缩放后的估计信号，通过预言机方法计算掩码，从输入中生成各个源的复数频谱图估计值。给定输入频谱图 I，输出源频谱图的计算方式为 $m_i * I$，其中 m_i 是使用频谱图估计的 i^{th} 声源对应的掩码。

表1：**结果**。我们将估计的幅度频谱图用作输入复数频谱图上的掩码，以生成各个声源的复数频谱图 [33]。下表展示了采用理想二值掩蔽和类维纳掩蔽时的尺度不变信噪比[dB] 和尺度不变信噪比提升[dB] 数值。数值越高越好。请注意，自编码器与 AudioSlots 之间的性能差距极小，这表明我们的处理流程能够很好地学习分离语音。

	尺度不变信噪比	尺度不变信噪比提升	尺度不变信噪比	尺度不变信噪比提升
掩蔽类型	IBM	IBM	维纳	维纳氏
预言机	12.07	12.07	12.32	12.33
自编码器	10.19	10.20	10.15	10.15
AudioSlots	09.50	09.50	09.96	09.97
预言机（1秒）	13.15	13.16	13.46	13.47
AudioSlots（1 秒）	09.66	09.67	10.20	10.20

由神经网络预测。这些复数频谱图通过逆短时傅里叶变换（STFT）转换为波形，然后拼接在一起；为简化处理，匹配时采用与真实信号的最佳匹配。

训练：我们使用 Adam [32] 进行训练，共执行 300k 步，批量大小为 64，学习率为 2e-4，预热 2500 步，并采用余弦衰减计划。

4. 实验

我们使用 Libri2Mix [5] 数据集评估所提方法在语音分离任务上的性能。我们使用该数据集的无混响版本。数据集中的每个样本均以 16kHz 采样，时长为 10 秒。Libri2Mix 包含来自 LibriSpeech [34] 的男性和女性说话人的话语。该数据集的 train-360-clean 划分包含 364 小时的混合信号，且声源均为无放回抽取。

如上所述，我们使用掩蔽，结合输入频谱图和网络预测，来估计各个源的复数频谱图。可使用多种掩蔽函数 [33]。在实验中，我们使用理想二值掩蔽（IBM） [35] 和类似维纳滤波器的掩蔽函数，其定义如下：

$$\text{IBM: } m_i = \begin{cases} 1 & i = \text{argmax}(m) \\ 0 & otherwise \end{cases}$$

$$\text{“Wiener like”: } m_i = \frac{(m_i)^2}{\Sigma_{i=1}^{n}(m_i)^2}$$

指标：我们使用尺度不变信噪比（SI-SNR） [36] 和尺度不变信噪比提升（SI-SNRi）来衡量分离性能。令 y 表示目标信号，$\hat{y}$ 表示通过我们的方法获得的估计信号。那么，SI-SNR 通过对目标信号进行重新缩放，衡量y 与$\hat{y}$ 在任意尺度下的保真度：

$$\text{SI-SNR}(y, \hat{y}) = 10 \log_{10} \frac{||\alpha y||^2}{||\alpha y - \hat{y}||^2}$$

其中 $\alpha = \text{argmin}_a ||ay - \hat{y}||^2 = y^T\hat{y}/||y||^2$。尺度不变信噪比提升是每个声源经处理后的估计信号的尺度不变信噪比，与将输入混合信号作为每个声源的估计信号时所得尺度不变信噪比之间的差值。评估时，我们首先匹配目标信号和估计信号，使尺度不变信噪比最大化，然后对所得的尺度不变信噪比和尺度不变信噪比提升分数取平均值。

结果：表1中，我们将本方法的性能与本方法的自编码器变体（以真实参考声源作为输入）以及使用（预处理后的）真实信号进行分离所取得的性能进行比较。使用真实信号计算出的指标代表了采用我们的（有损）预处理所能达到的最大值。在自编码器变体中，我们训练网络，使其使用 $n = 1$ 重建各个源信号。随后，我们将各个估计信号用作混合音频复数频谱图上的掩码。由于频谱图包含高频特征，这有助于我们了解该架构忠实表征这些特征的能力。我们还通过在预处理步骤中将裁剪长度增加到 1 秒进行了消融实验。

自编码器变体与分离模型之间的性能差异仅为 0.18 ± 0.01。这表明，我们的方法 AudioSlots 能够很好地学习语音分离，其性能已非常接近基线方法；后者以完全分离的声源作为输入。

不过，无论是我们的模型还是整体处理流程，都仍有很大的改进空间：以往基于掩蔽的方法 [11, 12, 13]已经能够以令人印象深刻的程度解决 Libri2Mix 说话人分离问题，其取得的尺度不变信噪比数值显著高于我们在此报告的结果。最近，基于扩散的方法也展现出了具有竞争力的性能 [37]。造成这一差距的部分原因在于我们的有损预处理流程：例如，对预先分块的音频计算短时傅里叶变换（STFT）（为简化处理而采用）会引入边界伪影，甚至使我们的真实标签尺度不变信噪比得分低于其他模型能够达到的水平。为简化处理，我们还将所有音频信号补零至相同长度。此外，由于我们从生成的频谱图中创建掩码，因此也受到基于掩蔽方法的局限，例如，被掩蔽的频谱内容无法重新生成。

局限性：我们的实验比较凸显了本方法的主要局限，即重建保真度：自编码器基线和我们的 AudioSlots 模型都使用潜在瓶颈表示对信号进行编码，因此往往会舍弃某些高频细节（以及其他细节）。这一点在图 2 中也能直观地看出。比较结果还表明，裁剪长度会影响分离性能：虽然这在一定程度上可以归因于我们简单的裁剪方法所导致的边界伪影，但也说明 AudioSlot 分离和重建音频频谱图的能力对分块长度较为敏感。

讨论：我们的结果表明，使用以槽位为中心的生成模型来处理音频分离具有良好前景；该模型使用一组特定于源的潜变量来表示音频。这与早期直接对输入音频进行操作的方法有显著不同，例如采用基于掩蔽的方法 [20]。学习特定于源的潜变量还有一个好处，即这些分解后的潜变量很可能不仅可用于生成，也可用于识别任务，这类似于以对象为中心的计算机视觉模型中的槽位如何作为目标检测和分割的基础。

我们乐观地认为，未来的工作可以克服当前方法的局限性：

• 为解决重建保真度问题（高频特征重建模糊），可以考虑放弃确定性前馈解码器，改用自回归解码方法（如 AudioLM [38], ），或迭代式扩散解码器（如 [39], ），这很可能有助于实现高保真度生成。

• 目前，AudioSlots 假设监督信号以真实声源的形式提供。若能扩展至直接使用原始

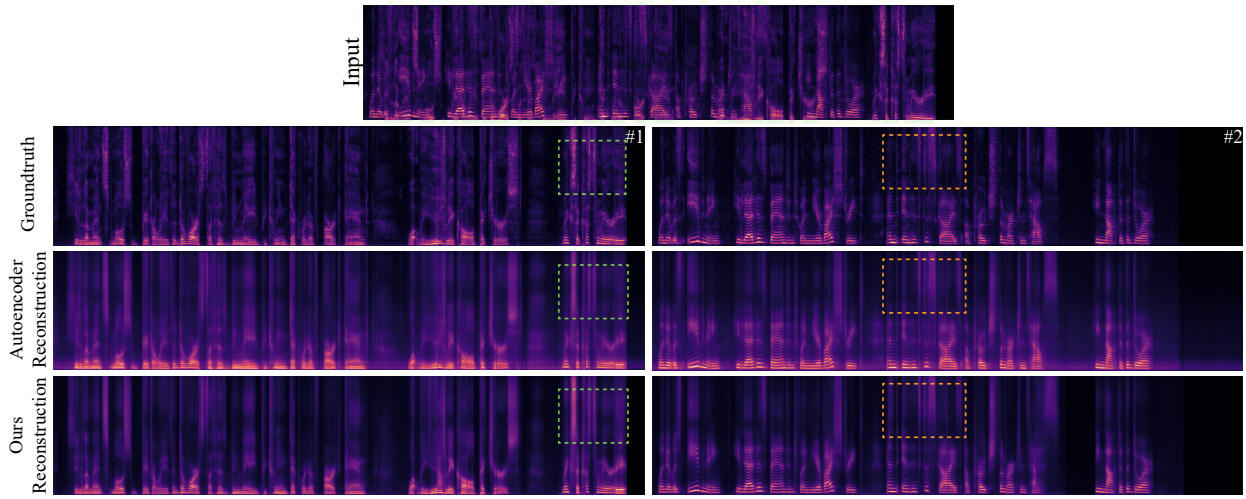

图2：真实值中各个声源的源频谱图、自动编码器估计结果与 AudioSlots（我们的方法）估计结果的绝对值对比。输入频谱图（**顶部**）是混合音频，其余各行显示各个声源的频谱图。输入频谱图和真实值频谱图均按照第 3 节所述步骤进行预处理。请注意，我们的方法能够较好地重建谐波，但在估计高频特征方面仍有困难（见高亮示例区域）。

混合音频进行完全无监督训练，将会很有趣。

为此，我们尝试将 AudioSlots 中的 Transformer 替换为 Slot Attention [3] 模块。该模块具有倾向于分解的归纳偏置，因此能够在视觉 场景分解的背景下进行无监督训练。然而，在初步实验中，我们发现，这种归纳偏置可能不足以完全以无监督的方式分解音频频谱图。不过，在初步实验中，采用 Slot Attention 模块的有监督版本 AudioSlots 表现与采用 Transformer 模块的版本相近，这表明进一步探索这一方向仍然是值得期待的未来研究工作。

- 我们认为，可以通过采用模型的序列扩展来克服单独处理各个分块所带来的局限性（这种处理方式需要事后拼接）：将前一时间步的槽位用作下一时间步的初始化，就像 Slot Attention for Video [40] 中那样。我们将此留待未来工作。

5. 结论

我们提出了 AudioSlots，一种以槽位为中心、用于音频频谱图的生成式架构。我们通过概念验证表明，AudioSlots 有望利用结构化生成模型解决音频源分离任务。尽管当前版本的 AudioSlots 存在若干局限，包括高频特征的重建保真度较低，以及需要以分离的音频源作为监督信号，但我们仍对此持乐观态度，并概述了若干未来工作的可能方向。

6. 参考文献

[1] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko，“使用变换器进行端到端目标检测”，载于*ECCV*，2020年。[2] Klaus Greff、Rapha¨el Lopez Kaufman、Rishabh Kabra、Nick Watters、Christopher Burgess、Daniel Zoran、Loic Matthey、Matthew Botvinick 和 Alexander Lerchner，“多目标采用迭代变分推断进行表征学习，”发表于*ICML*，2019年。[3] Francesco Locatello、Dirk Weissenborn、Thomas Unterthiner、Aravindh Mahendran、Georg Heigold、Jakob Uszkoreit、Alexey Dosovitskiy 和 Thomas Kipf，“基于 Slot Attention 的以对象为中心的学习”，发表于*NeurIPS*，2020年。[4] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin，“注意力就是你所需要的一切”，发表于*NeurIPS*，2017年。[5] Joris Cosentino、Manuel Pariente、Samuele Cornell、Antoine Deleforge 和 Emmanuel Vincent，“Librimix：一个用于通用语音分离的开源数据集”，*arXiv* 预印本 *arXiv:2005.11262*，2020年。[6] John R Hershey、Zhuo Chen、Jonathan Le Roux 和 Shinji Watanabe，“深度聚类：用于分割与分离的判别式嵌入”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2016年。[7] Yusuf Isik、Jonathan Le Roux、Zhuo Chen、Shinji Watanabe 和 John R Hershey，“使用深度聚类进行单通道多说话人分离”，*arXiv* 预印本 *arXiv:1607 02173*，2016年。。[8] Morten Kolbæk、Dong Yu、Zheng-Hua Tan 和 Jesper Jensen，“采用深度循环神经网络的语句级置换不变训练进行多说话人语音分离”，*IEEE/ACM* 音频、语音与语言处理汇刊，2017年。[9] Yi Luo 和 Nima Mesgarani，“Conv-TasNet：突破语音分离的理想时频幅度掩蔽”，*IEEE/ACM* 音频、语音与语言处理汇刊，2019年。[10] Ethan Manilow、Gordon Wichern、Prem Seetharaman 和 Jonathan Le Roux，“为音乐源分离开辟新路：用于研究训练数据质量和数量影响的 Slakh 数据集”，发表于*IEEE*音频与声学信号处理应用研讨会论文集 *(WASPAA)*，2019年。

[11] Yi Luo、Zhuo Chen 和 Takuya Yoshioka，“双路径循环神经网络：用于时域单通道语音分离的高效长序列建模”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2020年。[12]Cem Subakan、Mirco Ravanelli、Samuele Cornell、Mirko Bronzi 和 Jianyuan Zhong，“语音分离只需注意力机制”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2021年。[13] Kai Li、Runxuan Yang 和 Xiaolin Hu，“一种采用自顶向下注意力的高效编码器—解码器语音分离架构”，*arXiv* 预印本 *arXiv:2209.15200*，2022年。[14] Marco Tagliasacchi、Yunpeng Li、Karolis Misiunas 和 Dominik Roblek，“SEANet：一种多模态语音增强网络”，发表于*Interspeech* 论文集，2020年。[15] Zhong-Qiu Wang、Samuele Cornell、Shukjae Choi、Younglo Lee、Byeong-Yeol Kim 和 Shinji Watanabe，“TF-GridNet：让时频域模型在单耳说话人分离中重焕活力”，*arXiv* 预印本 *arXiv:2209 03952*，2022年。。[16] Felix Weninger、John R Hershey、Jonathan Le Roux 和 Bj¨orn Schuller，“用于单通道语音分离的判别式训练循环神经网络”，发表于*2014*年 *IEEE*全球信号与信息处理大会 *(GlobalSIP)*。IEEE，2014年，第577–581页。[17] Yuma Koizumi、Shigeki Karita、Scott Wisdom、Hakan Erdogan、John R Hershey、Llion J ones 和 Michiel Bacchiani，“DF-Conformer：采用线性复杂度自注意力、融合 Conv-TasNet 与 Conformer 的语音增强集成架构”，发表于*2021*年 *IEEE*音频与声学信号处理应用研讨会 *(WASPAA)*。IEEE，2021年，第161–165页。[18] Zhuo Chen、Yi Luo 和 Nima Mesgarani，“用于单麦克风说话人分离的深度吸引子网络”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2017年。[19]Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng，“NeRF：将场景表示为用于视图合成的神经辐射场”，发表于*ECCV*，2020年。[20] Scott Wisdom、Efthymios Tzinis、Hakan Erdogan、Ron Weiss、Kevin Wilson 和 John Hershey，“使用混合不变训练进行无监督声音分离”，发表于*NeurIPS*，2020年。[21] Oriol Vinyals、Samy Bengio 和 Manjunath Kudlur，“顺序很重要：集合的序列到序列”，*arXiv* 预印本 *arXiv:1511.06391*，2015年。[22] Manzil Zaheer、Satwik Kottur、Siamak Ravanbakhsh、Barnabas Poczos、Russ R Salakhutdinov 和 Alexander J Smola，“深度集合”，发表于*NeurIPS*，2017年。[23] Juho Lee、Yoonho Lee、Jungtaek Kim、Adam Kosiorek、Seungjin Choi 和 Yee Whye Teh，“集合变换器：基于注意力的置换不变神经网络框架”，发表于*ICML*，2019年。[24] Yan Zhang、Jonathon Hare 和 Adam Prugel-Bennett，“深度集合预测网络”，发表于*NeurIPS*，2019年。[25] Adam R Kosiorek、Hyunjik Ki m 和 Danilo J Rezende，“使用变换器进行条件集合生成”，*arXiv* 预印本 *arXiv:2006.16841*，2020年。[26] Yi Zhou、Hui Zhang、Hana Lee、Shuyang Sun、Pingjun Li、Yangguang Zhu、ByungIn Yoo、Xiaojuan Qi 和 Jae-Joon Han，“Slot-vps：面向视频全景分割的以对象为中心的表征学习”，发表于*CVPR*，2022年。[27] Pradyumna Reddy、Paul Guerrero 和 Niloy J Mitra，“概念搜索：通过直接优化发现视觉概念”，*arXiv* 预印本 *arXiv:2210.14808*，2022年。[28] Hu Xu、Juncheng Li、Alexei Baevski、Michael Auli、Wojciech Galuba、Florian Metze、Christoph Feichtenhofer 等，“能够聆听的掩码自编码器”，*arXiv* 预印本 *arXiv:2207.06405*，2022年。[29] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun，“用于图像识别的深度残差学习”，发表于*CVPR*，2016年。[30] Nicholas Watters、Loic Matthey、Christopher P Burgess 和 Alexander Lerchner，“空间广播解码器：一种用于在变分自编码器中学习解耦表征的简单架构”，*arXiv* 预印本 *arXiv:1901.07017*，2019年。[31] Matthew Tancik、Pratul Srinivasan、Ben Mildenhall、Sara Fridovich-Keil、Nithin Raghavan、Utkarsh Singhal、Ravi Ramamoorthi、Jonathan Barron 和 Ren Ng，“傅里叶特征使网络能够在低维域中学习高频函数”，发表于*NeurIPS*，2020年。[32] Diederik P Kingma 和 Jimmy Ba，“Adam：一种随机优化方法”，*arXiv* 预印本 *arXiv:1412 6980*，2014年。。[33]Hakan Erdogan、John R. Hershey、Shinji Watanabe 和 Jonathan Le Roux，“使用深度循环神经网络进行相位敏感且由识别增强的语音分离”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2015年。[34] Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur，“LibriSpeech：基于公有领域有声读物的自动语音识别语料库”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2015年。[35] Yipeng Li 和 DeLiang Wang，“论理想二值时频掩码的最优性”，《*Speech Communication*》，2009年。[36] Jonathan Le Roux、Scott Wisdom、Hakan Erdogan 和 John R Hershey，“信号失真比——半生不熟还是恰到好处？”，发表于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2019年。[37] Shahar Lutati、Eliya Nachmani 和 Lior Wolf，“分离与扩散：使用预训练扩散模型改进源分离”，*arXiv* 预印本 *arXiv:2301.10752*，2023年。[38] Zal´an Borsos、Rapha¨el Marinier、Damien Vincent、Eugene Kharitonov、Olivier Pietquin、Matt Sharifi、Olivier Teboul、David Grangier、Marco Tagliasacchi 和 Neil Zeghidour，“AudioLM：一种用于音频生成的语言建模方法”，*arXiv* 预印本 *arXiv:2209.03143*，2022年。[39] Curtis Hawthorne、Ian Simon、Adam Roberts、Neil Zeghidour、Josh Gardner、Ethan Manilow 和 Jesse Engel，“使用频谱图扩散进行多乐器音乐合成”，*arXiv* 预印本 *arXiv:2206.05408*，2022年。[40]Thomas Kipf、Gamaleldin F. Elsayed、Aravindh Mahendran、Austin Stone、Sara Sabour、Georg Heigold、Rico Jonschkowski、Alexey Dosovitskiy 和 Klaus Greff，“基于视频的条件对象中心学习”，发表于*ICLR*，2022年。