arXiv:2512.15532v1 [cs.SD] 17 Dec 2025

用于音乐源分离的条件化 UNet

Ken O'Hanlon[1,2],Basil Woods[2], Lin Wang[1], 和 Mark Sandler[1]

[1] 伦敦玛丽女王大学数字音乐中心[2] AudioStrip Ltd. 伦敦

摘要。本文提出一种用于音乐源分离（MSS）的条件化 UNet。音乐源分离通常由多输出神经网络执行，通常采用 U-Net，每个输出代表预定义乐器词汇表中的一个特定音轨。相比之下，条件化音乐源分离网络会同时接收与目标音轨相关的音频查询，以及要从中提取该音轨的信号。因此，无需使用严格限定的词汇表，从而能够实现更贴近现实的音乐源分离任务。由于缺乏合适的数据，这类任务中条件化方法的潜力一直未能充分展现；近期推出的 MoisesDb 数据集解决了这一问题。近期提出的方法 Banquet 使用该数据集，并在更大的词汇表上取得了令人鼓舞的结果。Banquet 使用频带分割循环神经网络，而非 UNet；其作者认为，UNet 不适用于条件化音乐源分离。我们对此提出反驳，并提出 QSCNet，这是一种新颖的条件化 UNet，它将神经网络条件化组件融入用于音乐源分离的稀疏压缩网络。我们发现，在若干音乐源分离任务中，QSCNet 的信噪比比 Banquet 高出 1 dB 以上，同时参数数量不到后者的一半。

1 引言

音乐源分离（MSS）旨在将代表一首歌曲的信号分离成多个不同的信号，每个信号包含歌曲中某种乐器的音轨。此前，音乐源分离被认为是一项极其困难的任务，除了人声分离等专门场景外，几乎没有取得什么成功 [2]。深度学习应用于音乐源分离后，系统性能得到持续提升[27] [9] [7] [3] [8] [24] [19] [18] [28], ，尤其是在四音轨任务 中，系统尝试分离人声、贝斯 和鼓 音轨，以及一个用于涵盖其他内容的其他 类别。分离通常由具有多个输出头的神经网络执行，每个输出头对应一种乐器 [7] [28], ；也可以为每种乐器分别训练独立的神经网络 [27] [3] [19]。这类神经网络大多采用基于 UNet 的架构 [9] [7] [3] [8] [24] [28] ，通过在对应的解码器和编码器模块之间建立跳跃连接，增强编码器/解码器网络。一个值得注意的例外是频带分割网络 [19] [18] ： 它们在单个模块中对互不重叠的频谱图频带进行编码，并以类似方式进行解码。频带分割循环神经网络（ BSRNN） [19] 与其他近期表现出色的网络有相似之处，例如在编码器与解码器之间的颈部采用双路径方法[3] [28] [24], ，以及学习复数掩码并将其应用于输入频谱图，以计算音轨近似值 [28] [3]。

四音轨任务之所以占据主导地位，一个原因是存在专为该任务构建的数据集 MusDb [22], 。有时，这项任务会扩展加入额外的分轨类别，例如吉他 [9], 或吉他和钢琴 [24]。然而，这些尝试需要私有数据，且仍然使用固定词汇表；此外，有些分轨类别也存在歧义 [24]。另一种方案是采用条件化音乐源分离网络 [29] [1] [26] [17] [5] [15] [30] ，通过消除网络输出中的硬性乐器类别，有望避免这些问题。在这类条件化网络中，输入信号会伴随一个音频查询，指明要分离的音轨。通常，系统会从音频查询中提取嵌入表示，并通过特征线性调制器（FiLM）[21] 层将其输入分离网络。FiLM层位于网络中的某个位置，用于调制通道激活，以突出某些特征；在此情形下，这些特征与特定音轨的信号相关。

与更普遍的音乐源分离问题类似，U-Net架构主要用于条件化音乐源分离。迄今为止，这些条件化U-Net大多采用MusDb [22] 作为主要数据集 [5] [15] [1] [29] [4] ，而URMP数据集 [16] 也曾结合类似架构使用 [26] [17]。MusDb限制了条件化音乐源分离的可能性，因为它只提供少量四种以活跃音轨为主的分轨类别。URMP确实提供了更丰富的音轨词汇表，但数据集规模较小，且录制方式较为单一。近期推出的数据集MoisesDb [20] 包含多轨音轨，并定义了一个音轨层级，分为11种不同的音轨类别，每种类别又包含若干子类别，例如吉他类别就包含 { 清音电吉他、失真电吉他、原声吉他}等子类别。MoisesDb具有如此丰富的本体，因此能够支持四音轨分离之外的新任务。Banquet [30], 是基于BSRNN [19] 的条件化音乐源分离神经网络，也是首批利用这一新资源的论文之一，并在更大的词汇表上取得了令人鼓舞的结果。在Banquet中，研究者使用最先进的音乐乐器识别网络PASST [14], 来提取查询嵌入，并将查询输入置于解码器之前的FiLM层[21] ，与 [29]类似。这与早期神经网络中采用的多种位置不同：每个解码器块中 [5], 、每个编码器块中 [1] [26],，以及每个卷积层中 [4] [13]。Banquet对BSRNN的改造基于这样一种观点：U-Net不适用于条件化音乐源分离，因为它们存在信息流问题 [30] [31], ；而BSRNN的频带拆分整体式编码器和解码器层则避免了这一问题。

本文认为，关于 UNet 中信息流不佳的论断可能并不成立，因为 UNet 中的跳跃连接应当有助于信息流动。事实上，迄今为止，大多数多输出音乐源分离网络都是 U-Net。因此，我们提出了 Query-SCNet（QSCNet），这是一种稀疏压缩网络（SCNet） [28], UNet架构的条件化变体，在四轨任务上的表现与 BSRNN 相近。我们在 Banquet论文 [30], 中概述的一些任务上，于 MoisesDb 上取得了更优的音乐源分离结果，尤其是在六声部问题上，信噪比提升了 1.6dB，改进幅度非常显著。同时，我们观察到，QSCNet 所需的参数量仅约为 Banquet 的 40%。

2 使用 UNet 进行音乐源分离

音乐源分离旨在将一条音乐音轨分离为若干组成音轨，每条音轨均包含某种乐器或乐器类别的信号。考虑一个音乐信号，$\mathbf{y} \in \mathbb{R}^{C\times N}$，其具有 C 个声道和 N 个采样点，以及一组构成音轨词汇表的乐器$\mathcal{I}$，音乐源分离问题可以定义为寻找源集合 $\{\mathbf{s}_i \in \mathbb{R}^{C\times N}\}$，使得:

$$\mathbf{y} \approx \sum_{i\in\mathcal{I}} \mathbf{s_i}.$$

这是一个难题，因为对于给定的曲目，$\mathcal{I}$ 中的音轨成员构成及数量可能未知，而且音轨数量通常多于声道数；立体声录音有 2 个声道。迄今为止，大多数音乐源分离工作通常会预先定义所用乐器的集合，例如 $\mathcal{I}^4 = \{bass, vocals, drums, others\}$，

并将神经网络 $\mathcal{N}$ 应用于信号，直接得到多个对应的输出

$$\mathcal{N}(\mathbf{y}) \longrightarrow \{\mathbf{s_i}\}_{\mathbf{i}\in\mathcal{I}} \tag{1}$$

不过，有些情况下会为每种乐器分别训练一个神经网络

$$\mathcal{N}_i(\mathbf{y}) \longrightarrow \mathbf{s_i}. \tag{2}$$

UNet [23] 是音乐源分离问题中常用的一种架构，通常应用于复数频谱图: $\mathbf{X} \in \mathbb{C}^{F\times T} = \text{STFT(y)}$。U-Net 可以视为由几个常见的子网络组成：编码器 $\mathcal{N}^{Enc}$、解码器$\mathcal{N}^{Dec}$，以及连接这两个子网络的颈部 $\mathcal{N}^{Neck}$。虽然最初的 UNet 是全卷积的，但音乐源分离中使用的变体通常将 $\mathcal{N}^{Neck}$设计为基于循环神经网络的[8] [28], 模块或基于 Transformer 的 [24] 模块。

与标准编码器／解码器架构类似，UNet 中的编码器由若干个 L 个连续模块组成，这些模块会降低特征维度$\mathcal{N}^{Enc} = (\mathcal{N}^{e_l})_{l\in\mathcal{L}}$ 其中 $\mathcal{L} = \{1, .., L\}$，而解码器则类似地由若干个模块组成 $\mathcal{N}^{Dec} = (\mathcal{N}^{d_l})_{l\in\mathcal{L}}$，用于提高特征维度。在时频域音乐源分离中，UNet 的编码器和解码器模块通常只改变频率方向上的维度，而时间维度的大小在整个神经网络中保持不变，不过也有例外 [3]。U-Net 的一个标志性特征是，具有相似表示维度的编码器和解码器模块通过跳跃连接相连。给定$\mathbf{e_l} = \mathcal{N}^{e_l}(\mathbf{e_{l-1}})$ 表示第 l 个编码模块的输出，则第 $(L-l)$ 个解码模块的操作可表示为 $\mathbf{d_l} = \mathcal{N}^{d_l}(\mathbf{d_{l-1}}, \mathbf{e_{L-l}})$。

基于掩码的音乐源分离 UNet 这一特例可视为以下操作序列

$$\begin{aligned}
\mathbf{E} &= \mathcal{N}^{Enc}(\mathbf{X}) \\
\mathbf{N} &= \mathcal{N}^{Neck}(\mathbf{E}) \\
\{\mathbf{M_i}\} &= \mathcal{N}^{Dec}(\mathbf{N}, \{\mathbf{e_l}\}_{l\in\mathcal{L}}) \\
\{\mathbf{s_i}\} &= \{\text{ISTFT}(\mathbf{M_i} \otimes \mathbf{X})\}_{i\in\mathcal{I}}
\end{aligned} \tag{3}$$

其中 $\mathbf{M}_i$ 表示与 $\mathcal{I}$ 中的第 i 个音轨相关的复数掩码，$\mathbf{E} = \mathbf{e_L}$。

2.1 条件化音乐源分离

音乐源分离的另一种方法不需要固定的音轨词汇表 (1)，也不需要逐音轨网络的集成 (2)。音乐源分离的条件化方法使用一个神经网络处理所有可能的乐器，并将类别 [25] 的表示或音频查询 $\mathcal{Q}$与输入信号一同作为输入：

$$\mathcal{N}(\mathbf{y}, \mathcal{Q}_i) \longrightarrow \mathbf{s_i}. \tag{4}$$

这种方法也可称为基于查询的源分离。虽然 $\mathcal{Q}$ 可以是一个独热向量，其中每个维度表示某个给定音轨的活动情况 [25]，但将 $\mathcal{Q}$ 作为查询嵌入提供会更加灵活，因为它表示音频特征空间中的一个点或区域。查询嵌入通常由一个独立的神经网络输出，例如由一个经过音乐乐器识别训练的神经网络输出，并且通常取自该神经网络倒数第二层的激活值。

采用特征级线性调制（FiLM） [21] 模块，根据查询嵌入在神经网络的某个位置对网络激活值 $\mathbf{P} \in \mathbf{R}^{C \times F' \times T'}$ 进行条件化， $\mathcal{Q} \in \mathbb{R}^q$

$$\mathbf{P} \longleftarrow \mathcal{N}^{FilM}(\mathcal{Q}, \mathbf{P}). \tag{5}$$

FiLM 模块 $\mathcal{N}^{FiLM}$ 会学习一个仿射变换函数，其参数为 $\mathbf{o}^{\gamma}$,$\mathbf{o}^{\beta}$ ，这些参数通常由响应向量查询输入 $\mathcal{Q}$的小型神经模块生成：

$$\gamma = \mathbf{o}_{\gamma}(\mathcal{Q}); \qquad \beta = \mathbf{o}_{\beta}(\mathcal{Q}) \tag{6}$$

将所得向量 $\gamma, \beta \in \mathbb{R}^C$ 应用于输入表示

$$\mathbf{P}_{c,f,t} \longleftarrow \gamma_c \mathbf{P}_{c,f,t} + \beta_c. \tag{7}$$

除基于查询的调制外，条件化音乐源分离在掩码生成和输出阶段也不同于多音轨方法（3）。此时，掩码集合中只有一个成员 $\mathbf{M}$，因此仅产生一个输出信号 $\mathbf{s}$。不过，仍然可以通过将多个查询与一个信号一起输入，并以批次形式进行 FiLM 后处理，在一个批次中完成多音轨分离。这样，由于信号编码只需执行一次，推理时间的计算负载可能会降低。

3 提出的方法

我们提出了一种用于音乐源分离（MSS）的条件化 UNet，称为 Query-SCNet（QSCNet）。QSC-Net 是对稀疏压缩网络 [28] (SCNet) 的改造，并为其增添了条件化能力。我们选择采用 SCNet 的理由之一，是它与 BSRNN 存在一些相似之处，而 Banquet 正是基于 BSRNN 构建的。在四音轨音乐源分离任务中，BSRNN 和 SCNet 的表现被认为相近 [28]，且二者都具备现代音乐源分离网络的特征，例如双路径循环神经网络和复数掩码学习。选择 SCNet 的另一个考量是，与其他先进的音乐源分离网络相比，它的规模相对较小 [28]。

3.1 SCNet

原论文的作者 [28] 并未将 SCNet 描述为 UNet。相反，他们着重说明了一个额外的融合层，该层接收与 UNet解码器模块相似的输入（$\mathbf{e_l}$，$\mathbf{d_{L-l}}$），并将输出传递给解码器。不过，我们在此指出，只需将论文 [28]中定义的融合层和解码器层合并，便可将 SCNet 表述为一种 UNet变体（3）。据此来看，SCNet 与其他 U-Net 的主要区别在于：它在编码器和解码器中采用了新颖的分带下采样和上采样模块，并使用了一种新颖的双路径循环神经网络。我们在此简要介绍这些特征，因为 QSCNet 自然继承了它们。

在 SCNet 中，立体声复数谱图首先被汇集为一个包含 4 个声道的张量。在该张量的每个时频点上，4 个声道分别表示两个立体声声道中对应复数短时傅里叶变换的实部和虚部系数 [28]。在编码器的每个后续块中，输入都会沿频率维度按照固定比率进行粗粒度分带。由此得到 3 个独立的分带张量，随后分别对其应用不同的下采样比率和卷积处理策略，最后在编码器块的输出处重新汇集。尽管分带过程采用了固定比率，但需要注意的是，这并不意味着整个解码器中都保持一致的频带划分，因为在第一个下采样层中，分带仅应用于具有线性频率尺度的维度。解码器中则采用相应的重组策略。

SCNet 的颈部由 6 个交替处理的双路径双向长短期记忆网络组成。通常，在音乐源分离（MSS） [3] [19] 中，双路径循环神经网络是指先使用循环神经网络沿时间维度进行处理，再以类似方式沿频率维度进行处理。SCNet 的颈部采用了一种新颖的变体：交替的双路径循环神经网络在不同域上运行 [28]。第一个以及后续编号为奇数的双路径循环神经网络直接作用于潜在特征，而编号为偶数的双路径循环神经网络则作用于潜在特征的傅里叶表示。在奇数和偶数双路径循环神经网络之间，对特征图应用实数快速傅里叶变换，并在潜在特征的频域中进行处理。处理完成后，在偶数和奇数双路径循环神经网络之间应用逆实数快速傅里叶变换，将基于傅里叶的特征图转换回原始的潜在特征域 [28]。

3.2 条件化

QSCNet 使用基于嵌入的查询。与 [30]，类似，我们采用 PASST 神经网络 [14] 提取嵌入，从而保持 QSCNet 与 Banquet 之间的相似性。具体而言，我们使用了 PASST 的一个变体，该变体是在 OpenMic [10] 数据集上针对乐器识别任务训练的，并且作者免费提供了该模型 [14]。此版本的神经网络仅有 20 个输出，分别代表不同的乐器或乐器类别

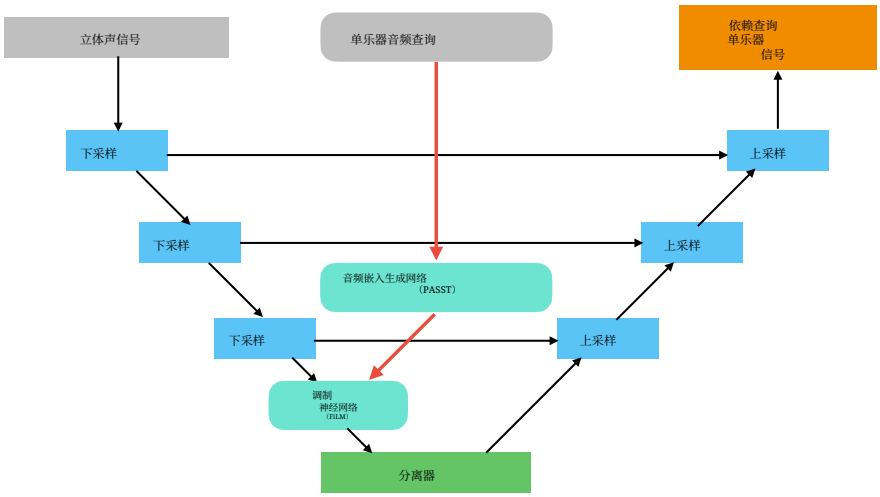


图1. QSCNet（或通用 UNet）的示意图，其中 L=3、一个 PASST 嵌入生成网络和一个 FiLM 调制网络集成在编码器末端。

同样，与 [30] 类似，我们移除最后一层，并将输出用作嵌入 $\mathcal{Q} \in \mathbb{R}^{768}$，输入为最长 10 秒的音频片段。

我们只使用了一个 FiLM 模块，将其置于编码器末端、双路径循环神经网络之前，如图 1 所示。换言之，在 (5) 中，我们设定 $\mathbf{P} = \mathbf{E}$，其中 $\mathbf{E}$ 是编码器的输出 (3)。虽然我们知道，这并非条件化音乐源分离中常用的 FiLM 位置，但我们认为此处应是进行条件化的最佳位置，因为这样可以在序列长期处理开始前定义乐器上下文。我们注意到，已有多项研究采用了不同的条件化位置，并在不同位置使用了多个 FiLM 模块。这里或许最值得关注的是，在 Banquet 中，条件器置于 $\mathcal{N}^{Neck}$ 末端，即解码器开始之前。

此处使用的 FiLM 模块为仿射变换 (7) 的参数$\mathbf{o}^{\gamma},\mathbf{o}^{\beta}$ (6)采用了类似的神经网络。这些神经网络由包含两个全连接层的多层感知机组成；第一层有 $q = 768$ 个输入和 $c = 128$ 个输出，后接指数线性单元激活函数 [6]；第二层分别有 $c = 128$ 个输入，输出为 γ 或 β (6)。

4 实验

我们进行了实验，以检验所提出的 QSCNet 的有效性。我们使用 6 个音轨的词汇表 $\mathcal{I}^6 = \{vocals, bass, drums, guitar, piano, others\}$训练 QSCNet，该词汇表与[24]中采用的词汇表类似。我们还提出了 SCNet 的六音轨变体（SCNet6），并使用相同的词汇表进行训练。 $\mathcal{I}^6$ 包含相同的

乐器音轨，与 [30],中采用的 Q:VDBGP 词汇表相同，因此可以直接与 Banquet 进行比较。不过， $\mathcal{I}^6$ 还包含 $others$ 类别，因此可以更直接地比较 QSCNet 和 SCNet6。

由于条件网络模型可以接受任意输入查询并返回相应输出 (4)，因此它具有一定的灵活性：与多声部网络不同，使用一种词汇表训练的模型可以使用另一种词汇表进行测试。基于这一点，我们还在 $\mathcal{I}^6$ 的扩展变体上进行了测试，其中包含一些更细分的音轨，例如 男声和 女声，而不只是较粗略的 人声。

4.1 数据

我们使用了 MoisesDb [20]„，其中包含 240 首带有多轨音频的歌曲，并采用 [30] 提出的训练/验证/测试集划分，分别包含 $144/48/48$ 首歌曲。MoisesDb 采用层级结构，包含 11 个独立的音轨类别，每个类别下还可能包含更细分的音轨子类别，总计 30 个。然而，与 MusDb 等四音轨数据集不同，其数据分布存在偏斜。 [20]。来自 $\mathcal{I}^4$ 的贝斯、鼓和人声等音轨几乎出现在每首歌曲中，而某些乐器类别并未出现在所有数据划分中 [30]。

我们首先使用词汇表 $\mathcal{I}^6$构建了训练数据集。对于此类别中的每个音轨，我们将其所有更细的子类别音轨合并，以形成单一音轨。随后，由所有未被分配到某一乐器类别的剩余音轨构成了 $others$ 分轨。验证集和测试集也以类似方式构建。

在训练时，我们采用了随机数据采样策略：通过混合为不同分轨随机选取的片段来生成训练片段。事实证明，这类由互不相关的音乐分轨构成的杂乱混合音频在训练用于音乐源分离的神经网络方面表现更优 [11]。我们使用时长为 10 秒的音频片段训练 QSCNet 和包含 6 个分轨的 SCNet。首先，通过在每条音轨中选取间隔 1 秒的 10 秒音频片段，为每个分轨整理出一组候选片段。在进行简单的静音检测后，每个候选分轨片段都会被纳入最终片段池；该检测会剔除其中超过 50% 的采样点为零的片段。

训练时，会为每个音轨类别随机选取一个音频片段，并对音轨片段数据进行一些数据增强。这包括声道翻转，即交换某种乐器的立体声声道；以及符号翻转，即反转信号中所有元素的符号。这两种翻转数据增强均以相同的概率随机启用。此外，还会进行增益增强：对每个乐器音轨片段分别施加一个随机增益，其取值范围为 (0.25,1.25)。这与 Demucs 网络 [24]以及 SCNet [28]中使用的数据增强设置类似。随后，将各种经过数据增强的音轨片段混合，形成训练混合音频。

对于条件化方法，还为每种乐器分别建立了一个查询池。在这里，这一过程只需对上述训练中使用的音频片段池进一步筛选即可完成。具体而言，仅将其中少于 20% 的采样点为零的音频片段用作查询。在条件化训练中，一个

查询乐器从 $\mathcal{I}^6$ 中选取，每种乐器被选中的概率相同。随后从该池中随机选取一个查询，并将其与增强音频混合信号一同输入神经网络。类似地，在验证和测试时，分别从验证集或测试集中每种乐器对应的查询池中随机选取一个查询。

按照上述音轨子类别汇集策略，为 $\mathcal{I}^6$ 构建了一个测试集。使用相同的数据和扩展词汇表构建了另一个测试集，称为 $\mathcal{I}^{6E}$，其中以更细分的音轨替换了人声、吉他和钢琴类别。具体而言，人声分为男声和女声，吉他分为清音电吉他、失真电吉他和原声吉他，而钢琴则以三角钢琴和电钢琴子类别表示。如果包含 $others$类别，则词汇表共包含 10 个音轨类别。

4.2 参数

神经网络输入所用的频谱图由采样率为 44.1 kHz 的立体声音频文件生成，窗口大小为 4096，重叠率为 75%。与 SCNet [28]一样，使用均方根能量（RMSE）作为损失函数。实验在单个 $A100$ 芯片上进行，内存为 80 Gb，批次大小为 8。部分初步实验采用了 4 和 16 的批次大小。我们发现，批次大小为 4 时训练效果有时较差，而批次大小为 16 时训练速度较慢，需要更多训练轮次。每个批次取 32000 个采样点，共得到 4000 个小批次。训练采用 Adam [12] 优化器，学习率设为 3×10^{-4}。每个模型训练 300 个训练轮次，并在每轮后进行验证。与 [24][28] 类似，每轮后都会维护指数移动平均。最终保留在验证集上表现最佳的模型或平均模型，作为训练完成的模型。

4.3 评估指标

针对每条音轨，计算第 i个乐器的信噪比

$$SNR_i = 10 \times \log_{10} \frac{\|\mathbf{y_i}\|_F^2}{\|\mathbf{y_i} - \mathbf{s_i}\|_F^2}$$

其中，$\mathbf{y_i}$ 和$\mathbf{s_i}$ 分别表示已知的音轨信号和近似的音轨信号。对于每种乐器，记录所有同类乐器音轨的信噪比中位数。这与 [30] 中使用的评估指标类似，可直接比较结果。

4.4 结果

$\mathcal{I}^6$ 上的结果如表 1 所示。表中将 QSCNet 与提出的 SCNet6 及其较大的变体 SCNet6(L) 进行了比较。SCNet6(L) 与 SCNet6 采用相同的架构，但每层的声道数增加了一倍。这里

算法	贝斯	人声	鼓	吉他	钢琴	五音轨平均值	其他
Banquet	11.0	8.0	9.5	3.3	2.5	6.9	-
QSCNet	11.9	9.8	11.7	5.7	3.4	8.5	1.3
HTDemucs	10.9	8.9	11.6	2.4	1.7	7.1	-
SCNet6	12.8	10.5	12.4	6.3	4.0	9.2	2.8
SCNet6(L)	13.5	12.2	13.4	7.0	4.6	10.1	3.4

表 1. 比较 6 音轨方法 HTDemucs6、SCNet6 和 SCNet6(L)，以及两种条件化方法 Banquet [30] 和 QSCNet 在 MoisesDb 测试集上进行音乐源分离的结果。按乐器给出的结果为信噪比中位数。

还将它们与 [30] 中给出的最先进结果进行了比较，包括六音轨版 HT-Demucs [24] 和经过条件化的频带分割网络 Banquet [30] 。由于未考虑其他 类别[30]，这两种方法都只针对 5 个音轨进行了评估。为便于简要比较，还记录了每种算法在 5 个乐器音轨上的平均分数（五音轨平均值）。

在多输出网络方面，SCNet6 的表现明显优于 HT- Demucs，领先 2.1 dB，且所有乐器的表现均有所提升。其中一些提升幅度很大，例如吉他的表现提高了 3.9 dB。考虑到 SCNet6 仅使用 MoisesDb 中的 144 首歌曲进行训练，而 HT-Demucs 使用包含 800 首歌曲的私有数据集进行训练 [24]，这一结果或许更令人瞩目。SCNet6(L) 进一步提升，五音轨平均值指标比 HTDemucs 高 3 dB。与标准 SCNet6 相比，SCNet6(L) 在人声和鼓类别上有显著提升。

考虑到条件网络，与 Banquet 相比，所提出的 QSCNet 在 Avg5 指标上以 1.6dB 的较大优势展现出更优性能。使用 QSCNet 时，所有乐器的表现均有所提升，其中鼓和吉他的提升尤为明显，二者相较于 Banquet 的提升都超过了 2dB。与 HTDemucs 的 6 音轨版本相比，QSCNet 的表现也有所提升，并达到了比 SCNet6 低约 0.7dB 的性能水平。值得注意的是，QSCNet 使用的参数远少于 SCNet6，因为 SCNet6 在独立掩码的构建中使用了大量参数。我们观察到，QSCNet 使用 10.2M 个参数；而 SCNet4 和 SCNet6 分别使用 20.4M 和 26.6M 个参数。QSCNet 也明显小于 Banquet，后者使用了 24.9M 个参数。

			人声		吉他			钢琴			
	鼓	贝斯	男性	女性	原声	清音	失真	三角钢琴	电声	平均值9	其他
Banquet	10.1	10.7	7.9	10.1	0.9	1.7	2.8	2.8	0.5	5.3	-
QSCNet	11.8	11.6	8.5	11.8	1.3	3.6	4.0	3.2	0.7	6.3	1.1

表2. 在扩展的 $\mathcal{I}^{6E}$ 词汇表下，Banquet[30] 与所提出的 QSCNet 在 MoisesDb 测试集上的音乐源分离结果对比。各乐器的结果以信噪比中位数表示。

扩展后的 $\mathcal{I}^6$ 上的结果如表2所示，其中 QSCNet 再次展现出优于 Banquet 的性能，这次平均提升了 1dB，且任何乐器都没有出现相对性能下降。在此情况下值得注意的是，Banquet 的训练使用了细粒度声部词汇 $\mathcal{I}^{6E}$ ，两种算法也都在该词汇上进行测试；而 QSCNet 则使用粗粒度声部版本 $\mathcal{I}^6$进行训练。

5 结论

我们提出了 QSCNet，这是 SCNet[28] 的一种条件化变体。实验结果表明，该方法显著优于 Banquet 模型；在粗粒度和细粒度声部上，针对 Q:VBDGP 问题的条件化音乐源分离均达到了最先进水平。我们认为，这验证了采用 SCNet 以及在 QSCNet 中选择 FiLM 模块位置的合理性。我们仅使用一个 FiLM 模块，并采用了远小于 Banquet 的神经网络，便取得了这些结果，因此我们认为，这有力地驳斥了 [30] 中关于 U-Net 不适用于条件化音乐源分离的负面论断。

我们还通过使用 MoisesDb 在六声部问题上训练 SCNet(6)，研究了六声部问题。SCNet6 和 SCNet6(L) 取得了最先进的结果，表现远超 HTDemucs6，尽管其训练所用的数据集小得多。有趣的是，就六个声部而言，QSCNet 的表现比 SCNet6 低 0.7 dB。这一点值得关注，因为 SCNet6 所需的参数数量是 QSCNet 的 250%，尽管两者使用的网络组件大小相近。我们认为，如果使用数量相近的参数，多音轨方法与条件化方法之间的性能差距或许还能进一步缩小。

尽管这里的性能提升相当显著，但我们确实认为，这可能并非完全归因于所选用的、接受条件控制的基础网络。未来的工作将对所提出的方法进行一些消融实验；一些未在此展示的初步结果表明，不同的神经网络和训练参数可能在一定程度上造成了性能差距。此外，我们还将研究在更大的词汇表上进行训练，例如 [30]，中的词汇表，并比较参数数量相近的条件神经网络与多输出神经网络的性能。

我们认为，这是一项尤其适时的贡献。条件化音乐源分离具备了以往音乐源分离所不具备的能力。这些能力很少得到充分展现，因为大多数关于条件化音乐源分离的研究都应用于 MusDb 等词汇表较小且固定的数据，而这种方法并不真正适用于此类数据。如今，随着 MoisesDb 的出现，已有可供研究的数据集，这应能推动更多关于条件化音乐源分离的研究。事实上，我们认为，鉴于其潜在能力，条件化音乐源分离应成为音乐源分离中最受关注的研究领域。我们认为，本文所呈现的工作是朝这一方向迈出的积极一步。

参考文献

1. Brocal, G.M., Peeters, G.：Conditioned-u-net：在 U-Net 中引入控制机制以实现多源分离。载于：第 20 届国际音乐信息检索学会会议论文集。第 159–165 页（2019） 2. Chan, T.S., Yeh, T.C., Fan, Z.C., Chen, H.W., Su, L., Yang, Y.H., Jang, R.：利用 i kala 数据集进行人声活动引导的歌声分离。载于：2015 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 718–722 页（2015） 3. Chen, J., Vekkot, S., Shukla, P.：基于轻量级深度学习框架的音乐源分离（DTTNet：双路径 TFC-TDF UNet）。载于：2024 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 656–660 页（2024） 4. Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., Dubnov, S.：通过基于查询的学习，从弱标注数据中实现零样本音频源分离。载于：AAAI（2022） 5. Choi, W., Kim, M., Chung, J., Jung, S.：LASAFT：用于条件源分离的潜在源注意频率变换。载于：2021 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 171–175 页（2021） 6. Clevert, D.A., Unterthiner, T., Hochreiter, S.：利用指数线性单元实现快速、准确的深度网络学习。载于：第 4 届国际学习表征会议，ICLR

2016 1511 07289 。

7 ´

http://arxiv.org/abs/ . . Defossez, A., Usunier, N., Bottou, L., Bach, F.R.：波形域中的音乐源分离。arXiv 预印本 arXiv:1911.13254（2019） 8. D´efossez, A.：混合频谱图与波形源分离。arXiv 预印本 arXiv:2111.03600（2021） 9. Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.：Spleeter：一种采用预训练模型的快速高效音乐源分离工具。《开放源代码软件期刊》**5**(50), 2154（2020） 10. Humphrey, E.J., Durand, S., McFee, B.：OpenMIC-2018：用于多种乐器识别的开放数据集。载于：第 19 届国际音乐信息检索学会会议论文集（ISMIR）（2018） 11. Jeon, C.B., Wichern, G., Germain, F.G., Le Roux, J.：为何音乐源分离能从喧嚣声中受益？载于：2024 年 IEEE 声学、语音与信号处理国际会议研讨会（ICASSPW）。第 873–877 页（2024） 12. Kingma, D.P., Ba, J.：Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980（2017） 13. Kong, Q., Chen, K., Liu, H., Du, X., Berg-Kirkpatrick, T., Dubnov, S., Plumb- ley, M.D.：利用弱标注数据实现通用源分离。arXiv 预印本 arXiv:2305.07447（2023） 14. Koutini, K., Schl¨uter, J., Eghbal-zadeh, H., Widmer, G.：利用 Patchout 高效训练音频 Transformer。载于：Interspeech。第 2753–2757 页（2022） 15. Lee, J.H., Choi, H.S., Lee, K.：基于音频查询的音乐源分离。载于：第 20 届国际音乐信息检索学会会议论文集（2019） 16. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.：创建用于多模态音乐分析的多轨古典音乐演奏数据集：挑战、见解与应用。《IEEE 多媒体汇刊》**21**(2), 522–535（2019）

17. Lin, L., Kong, Q., Jiang, J., Xia, G.：用于零样本音乐源分离、转录与合成的统一模型。载于：第 22 届国际音乐信息检索会议论文集，ISMIR（2021）18. Lu, W.T., Wang, J.C., Kong, Q., Hung, Y.N.：采用频带拆分 RoPE Transformer 的音乐源分离。载于：2024 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 481–485 页（2024）19. Luo, Y., Yu, J.：采用频带拆分循环神经网络的音乐源分离。IEEE/ACM 音频、语音与语言处理汇刊 **31**，1893–1901（2023）20. Pereira, I., Ara´ujo, F., Korzeniowski, F., Vogl, R.：Moisesdb：一个超越 4 条音轨的源分离数据集。载于：第 24 届国际音乐信息检索学会会议论文集（ISMIR）（2023）21. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.：FiLM：利用通用条件层进行视觉推理。载于：AAAI（2018）22. Rafii, Z., Liutkus, A., St¨oter, F.R., Mimilakis, S.I., Bittner, R.：MUSDB18-HQ——MUSDB18 的未压缩版本（2019 年 8 月），https://doi.org/10 3338373 23 5281. /zenodo. . Ronneberger, O., Fischer, P., Brox, T.：U-Net：用于生物医学图像分割的卷积网络。载于：医学图像计算与计算机辅助干预——MICCAI 2015。第 234–241 页（2015）24. Rouard, S., Massa, F., D´efossez, A.：用于音乐源分离的混合 Transformer。载于：2023 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 1–5 页（2023）25. Seetharaman, P., Wichern, G., Venkataramani, S., Le Roux, J.：用于音乐源分离的类别条件嵌入。载于：2019 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 301–305 页（2019）26. Slizovskaia, O., Haro, G., G´omez, E.：针对乐器演奏的条件源分离。IEEE/ACM 音频、语音与语言处理汇刊 **29**，2083–2095（2021）27. St¨oter, F.R., Uhlich, S., Liutkus, A., Mitsufuji, Y.：Open-Unmix——音乐源分离的参考实现。《开放源代码软件期刊》**4**(41)，1667（2019）28. Tong, W., Zhu, J., Chen, J., Kang, S., Jiang, T., Li, Y., Wu, Z., Meng, H.：SCNet：用于音乐源分离的稀疏压缩网络。载于：2024 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 1276–1280 页（ 2024 ） 29. Wang, Y., Stoller, D., Bittner, R.M., Bello, J.P.：少样本音乐源分离。载于：2022 年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第 121–125 页（2022）30. Watcharasupat, K.N., Lerch, A.：一种适用于四条以上音轨音乐源分离、与音轨无关的单解码器系统。载于：第 25 届国际音乐信息检索学会会议论文集（2024）31. Watcharasupat, K.N., Lerch, A.：分离这个，以及它周围的所有内容：通过超椭球查询实现音乐源分离。arXiv 预印本 arXiv:2501.16171（2025）