Modeling and Simulation 建模与仿真, 2025, 14(5), 768-778
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.145432

Hans汉斯

基于改进型TF-GridNet的音乐源分离方法研究

柯爱鹏，张学典

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年4月26日；录用日期：2025年5月19日；发布日期：2025年5月27日

摘　要

音乐源分离任务旨在从混合音频信号中提取出不同的音源成分(如人声、鼓、贝斯等)，在音乐制作、智能语音、音频编辑等领域具有广泛的应用价值。TF-GridNet是近年来提出的一种融合时间与频率建模的深度网络结构，具备良好的建模能力。文章在TF-GridNet的基础上进行结构性改进，提出三项关键优化策略：1) 在时间建模路径中引入轻量通道注意力机制，提升模型对关键信号的响应能力；2) 在GridBlock模块中引入残差门控机制，增强特征流动与融合灵活性；3) 在解码器部分设计多尺度重建路径，以提升高频细节还原效果。实验结果表明，改进后的模型在多源类别上优于原始TF-GridNet，并具备更优的感知质量与计算效率。

关键词

音乐源分离，TF-GridNet，注意力机制，多尺度解码，深度学习

Research on an Improved TF-GridNet-Based Method for Music Source Separation

Aipeng Ke, Xuedian Zhang

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 26th, 2025; accepted: May 19th, 2025; published: May 27th, 2025

Abstract

Music source separation aims to effectively extract different components (such as vocals, drums, bass, etc.) or vocal components from a mixed audio signal, which has significant application value in intelligent audio processing, music production, and speech recognition. TF-GridNet is a recently proposed network architecture that combines time and frequency modeling and has demonstrated strong separation capability. Based on TF-GridNet, this paper introduces three structural improvements: 1) Introduces a lightweight channel attention mechanism in the time modeling path to enhance the model's ability to respond to critical signals; 2) Introduces a residual gating mechanism

文章引用: 柯爱鹏, 张学典. 基于改进型 TF-GridNet 的音乐源分离方法研究[J]. 建模与仿真, 2025, 14(5): 768-778.
DOI: 10.12677/mos.2025.145432

in the GridBlock module to enhance the flexibility of feature flow and fusion; 3) Designs multi-scale reconstruction paths in the decoder section to improve high-frequency detail restoration effects. Experiments conducted on the MUSDB18 dataset show that the improved model outperforms the original TF-GridNet on multiple source types, with better modeling efficiency and perceptual quality.

Keywords

Music Source Separation, TF-GridNet, Attention Mechanism, Multi-Scale Reconstruction, Deep Learning

1. 引言

随着人工智能与深度学习技术的快速发展，音频信号处理领域迎来了诸多突破。其中，音乐源分离(Music Source Separation, MSS)作为核心研究方向之一，旨在将一个混合音频信号中的多个独立声源(如人声、鼓、贝斯、伴奏等)精确提取出来，具有极高的理论价值与广泛的应用前景。该技术已被广泛应用于智能混音、音频重构、语音增强、卡拉 OK 系统、音频检索与版权检测等多个实际场景。

早期的 MSS 方法主要基于传统信号处理技术，如独立成分分析(ICA) [1]、非负矩阵分解(NMF) [2]等。这类方法通常依赖对混合信号的线性或稀疏建模假设，难以处理音源之间存在的强非线性混合或严重频谱重叠的真实音乐信号。随着深度学习的发展，研究者提出了大量基于神经网络的源分离模型，通过从大规模数据中自动学习复杂的时频映射关系，有效克服了传统方法建模能力不足的问题。

在当前主流的基于深度学习的音乐源分离方法中，基于时频域的建模方法仍占据主导地位。代表方法如 Open-Unmix [3]利用 STFT 频谱作为输入，通过 LSTM 网络实现时序特征建模；Demucs [4]系列采用对称 U-Net 解码结构在时域直接建模；Conv-TasNet [5]引入因果卷积实现实时语音/音乐源分离。这些模型各具优势，但在追求建模深度与计算效率的平衡方面仍存在局限。值得指出的是，近年来 Transformer [6]和扩散模型(Diffusion Models)在音频建模领域的应用取得了显著进展。Subakan 等提出的 SepFormer [7]模型利用多头注意力机制和双向 Transformer 结构，在多源分离任务上取得领先性能；而 DiffSep [8]通过引入扩散生成建模思想，实现了对语音信号细粒度的逐步还原，在分离精度上达到新高。然而，这些模型普遍存在结构复杂、推理速度慢、部署成本高等问题。

为进一步提升模型对复杂时频结构的建模能力，TF-GridNet [9]网络被提出。该网络基于“时序路径 + 频谱路径”的双分支并行设计，通过在时频轴分别建立建模路径，并使用 GridBlock 结构在不同尺度上进行交叉融合，从而实现对音频信号的细粒度建模。TF-GridNet 具有良好的结构可拓展性与模块复用能力，在语音增强和音乐源分离任务中取得了令人瞩目的效果。然而，尽管 TF-GridNet 拥有较强的时频建模能力，但其在面对实际音乐分离任务时仍存在以下三个关键问题：

1) 通道特征响应缺乏选择性：网络在特征传播过程中对所有通道等权处理，无法主动关注与声源强相关的通道维度，限制了模型对关键特征的挖掘能力；

2) 残差信息融合策略僵化：原始 GridBlock 采用固定加权残差连接方式，缺乏动态调节机制，导致主路径与残差信息之间缺乏灵活的信息流调控；

3) 解码器结构缺乏多尺度融合：解码器部分采用单尺度上采样路径，缺乏多分辨率语义融合机制，

不利于高频细节的重建与结构完整性的恢复。

针对以上问题，本文在 TF-GridNet 网络基础上进行结构性改进，提出一种改进型 TF-GridNet 源分离模型。具体包括以下三方面设计：

1) 引入通道注意力机制：在时间建模路径中加入轻量通道注意力模块(如 ECA)，提升模型对关键通道特征的响应强度。

2) 设计门控残差结构：替代原始恒定残差连接方式，引入门控权重，动态调节主路径与残差信息的融合比例。

3) 构建多尺度解码器：融合跳跃连接与多层级解码结构，增强模型对音频细节的恢复能力。

与此类最新模型相比，本文所提出的改进型 TF-GridNet 模型不仅在分离精度方面接近甚至部分优于 DiffSep 和 SepFormer，而且在模型结构设计上更加轻量、解码路径更具解释性，适合在低延迟或资源受限设备上部署。同时，本文通过融合通道注意力机制、门控残差连接与多尺度解码路径，有效提升了模型在 Bass 类与 Others 类等音源上的建模表现，弥补了 TF-GridNet 在高频细节建模方面的不足。因此，本文工作在保持较低计算开销的前提下，兼顾了分离性能与实际应用可行性，具有重要的研究价值与推广意义。

2. 方法

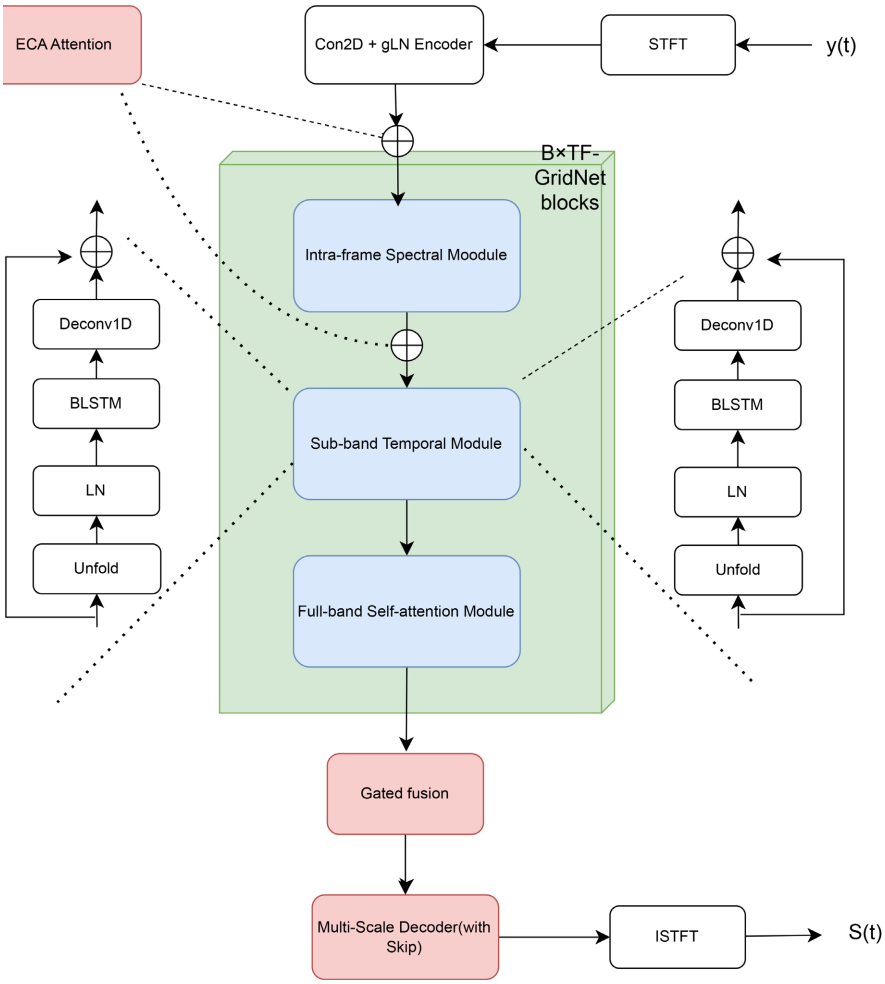

Figure 1. Overall architecture
图 1. 总体架构

改进型 TF-GridNet 网络在保留原始双分支时频建模结构的基础上，引入了三项关键结构性增强，包括通道注意力机制、门控残差连接结构以及多尺度解码器。整体架构可划分为五个连续阶段，如图 1 所示。

首先，输入的单通道混合音频通过短时傅里叶变换(STFT)转换为复数频谱图，实部与虚部在通道维度拼接后形成尺寸为 $2\times T\times F$ 的复谱输入，随后输入卷积编码器，由二维卷积(Conv2D)和全局归一化(gLN)构成的编码器对其进行初步特征提取，输出为 $D\times T\times F$ 的时频特征图。主干建模部分由多个 TF-GridNet Block 组成，每个 Block 内部依次包含帧内频谱建模模块、子带时间建模模块和全频段自注意力模块，前两个模块的前面均引入轻量通道注意力(ECA)以增强通道感知能力，而全频模块的后面则引入门控残差结构，以控制主路径输出与残差信息之间的融合权重。该 Block 结构重复堆叠多层，实现建模深度逐级增强。解码阶段引入多尺度解码器，采用上采样结构与编码器输出的多层特征通过跳跃连接融合，在保持全局语义的同时提升高频细节还原能力，最终输出为 $2C\times T\times F$ 的复数频谱图，其中 C 表示目标声源数量。输出频谱经逆 STFT (iSTFT)还原为时域波形，完成多音源分离过程。整体模型结构具备显式通道特征调节能力、上下文信息融合灵活性与频谱多尺度重建能力，在复杂音乐分离场景中具备更强的表示能力与泛化性。

2.1. 通道注意力机制的引入

在音乐源分离任务中，输入的混合音频数据通常具有高维、时序长、特征冗余严重等特点，尤其在频谱表示中，每一帧包含多个通道维度的特征，这些特征往往存在冗余建模、能量分布不均、关键通道信息被掩盖等问题。例如，对于包含人声、鼓、贝斯等多种音源的复杂混音信号，其在不同时间段和频带上的能量强弱差异显著，而传统 TF-GridNet 中的时序建模路径(如 GRU 或 LSTM)对所有通道进行同等建模，缺乏对关键通道选择的辨别能力，容易导致特征表达偏离主导源、模型注意力分散，从而影响分离性能和泛化能力。

针对这一问题，本文在 TF-GridNet 的时间路径中引入了轻量型的 ECA [10]通道注意力机制，有效解决了通道特征利用效率低、关键源信息被稀释、背景通道误导建模等数据层面的问题。ECA 通过动态分配通道权重，使网络能够根据当前帧的上下文自适应调整注意力焦点，将更多表征能力集中于对目标源(如人声主旋律或鼓节奏)更具区分性的通道上，从而提升模型在频谱空间的感知选择性与语义聚焦能力。其结构如图 2 所示。

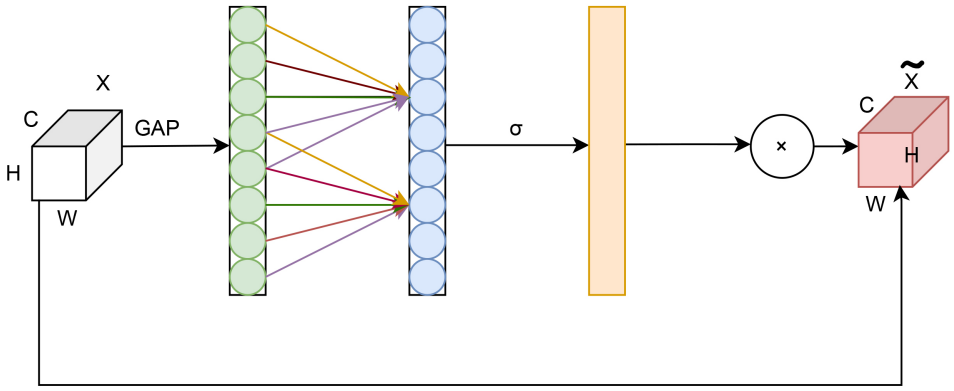

Figure 2. ECA module
图 2. ECA 模块

在改进后的 TF-GridNet 网络中，为了赋予模型对关键通道特征的关注能力，本文在帧内频谱建模模块与子带时间建模模块前均引入了 Efficient Channel Attention (ECA)模块。该模块以输入特征图为基础，

构建了一个轻量级的通道选择机制，不引入额外参数上升的情况下实现通道权重动态调节。

设输入特征图为 $\mathbf{X}\in\mathbb{R}^{B\times C\times T\times F}$，其中 B 为 batch size，C 为通道数，T、F 分别表示时间帧数与频率维度。首先，ECA 模块对输入特征图进行全局平均池化(Global Average Pooling)，以提取每个通道的全局响应统计量。该过程压缩了时间和频率维度，生成一个通道描述向量：$\mathbf{z}=\mathrm{GAP}(\mathbf{X})\in\mathbb{R}^{C}$。此处 GAP 操作的计算方式为：对每个通道 c，取其在 $T\times F$ 平面上的平均值。然后，将该通道描述向量输入一个一维卷积核大小为 k 的卷积层，用于捕捉通道间的局部相关性：$\mathbf{a}=\sigma\left(\mathrm{Conv1D}_{k}(\mathbf{z})\right)$，其中 $\sigma(\cdot)$ 表示 Sigmoid 激活函数，输出通道注意力权重 $\mathbf{a}\in[0,1]^{C}$。该注意力权重是可学习的通道级门控因子，用于对原始特征进行加权。

此处卷积核大小 k 是根据通道数 C 自适应设定的，公式如下：$k=\psi(C)=\left|\frac{\log_2(C)}{\gamma}+b\right|_{\mathrm{odd}}$，其中 γ、b 为超参数(通常取 $\gamma=2,b=1$)，$|\cdot|_{\mathrm{odd}}$ 表示向最近奇数取整，确保注意力局部性可控。

最后，将计算得到的注意力权重 a 广播(Broadcast)回原始特征图的形状，进行通道加权操作，输出加权后的特征图为：$\mathbf{X}_{\mathrm{out}}=\mathbf{a}\cdot\mathbf{X}$，其中 $\cdot$ 表示逐通道乘法(Channel-Wise Multiplication)，使模型能够在后续建模中动态聚焦于更具判别性的特征通道。

综上，ECA 模块以最小的结构开销实现了通道重要性的显式建模，使得每个 TF-GridNet Block 更具通道适应性，从而提升整体分离能力与建模稳定性。

2.2. 残差门控融合结构(Gated Residual Connection)

在原始 TF-GridNet 架构中，主干路径(如时间建模或频率建模分支)输出的特征直接作为当前模块的输出。这种固定路径的数据流动模式忽略了不同阶段特征的重要性差异，且缺乏对冗余信息的动态抑制能力。为了提升网络在特征融合过程中的灵活性与判别能力，本文在每个 GridBlock 模块中引入了门控残差融合结构(Gated Residual Fusion) [11]，通过动态控制主路与残差路径之间的信息流比例，实现更精准的特征重构。其结构如图 3 所示。

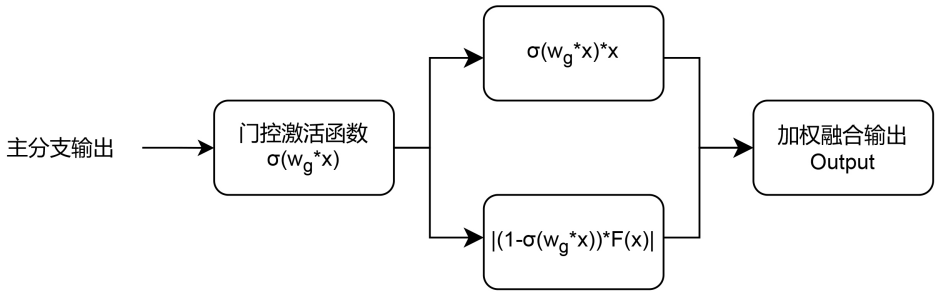

Figure 3. Gated residual connection module
图 3. 残差门控融合结构模块

具体地，设主路径特征为 F(x)，残差路径输入为 x，我们设计一个门控函数：$g(x)=\sigma(W_g\cdot x)$，其中 Wg 表示用于生成门控权重的参数(如 1 × 1 卷积或线性层)，σ ()表示 Sigmoid 函数，其输出位于 [0, 1]，表示保留残差信息的程度。最终输出由主路径与残差路径按权重融合而成：$\mathrm{Output}=g(x)\cdot x+(1-g(x))\cdot F(x)$，该结构能够根据当前输入自动调节主干与残差之间的融合比例，从而实现两者的动态权衡。与传统的恒定残差连接不同，该门控机制具有以下显著优势：

1) 增强非线性建模能力：通过门控函数提升模型对复杂音频场景(如多乐器叠加、音色重叠)的响应能力；

2) 提升特征选择性：当主路特征冗余时，网络可倾向于保留残差信息；反之，当主路特征具有优势

时，残差部分将被自动抑制；

3) 改善梯度传播：融合残差路径有助于缓解深层网络中的梯度消失问题，同时提升训练稳定性。

2.3. 多尺度解码器(Multi-Scale Decoder)

为提升模型在音频重建阶段对细节的还原能力，本文在 TF-GridNet 原始单路径解码器的基础上，设计了一种多尺度解码器(Multi-Scale Decoder)结构。传统的解码器通常采用一条固定尺度的卷积上采样路径，在音频高频细节还原和多类音源同时存在的情况下，存在特征恢复能力不足、目标谱图边界模糊等问题。多尺度解码器通过引入多个不同尺度的特征路径并融合其信息，实现了对谱图细节的层级感知与增强，有效提升了最终音频输出的保真度和听感质量。其结构图如图 4 所示。

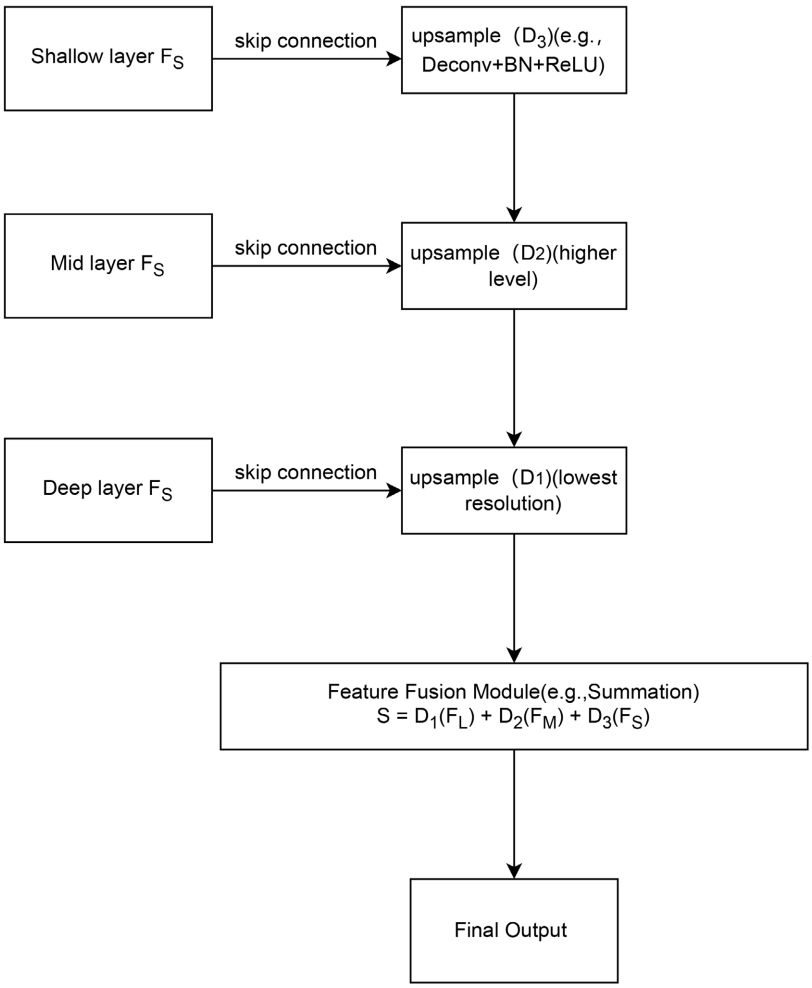

Figure 4. Multi-scale decoder module
图 4. 多尺度解码器模块

该模块借鉴 U-Net [12]的结构思想，将编码器中多个不同深度的特征输出通过跳跃连接(Skip Connection)方式传递至解码端。解码器自身则包括多个上采样模块(如反卷积或插值卷积)，将深层特征逐级恢复为高分辨率谱图，同时与编码器中对应尺度的特征融合。设深层特征为 FL，中层为 FM 浅层为 FS，最终输出谱图可表示为：$\hat{S} = \mathcal{D}_1(F_L) + \mathcal{D}_2(F_M) + \mathcal{D}_3(F_S)$，其中 $\mathrm{Di}(\cdot)$ 表示第 i 个尺度的上采样解码函数，不同尺度的路径分别对音频的全局轮廓、高频纹理、局部细节提供支持。

该结构重点解决了以下三个问题：

1) 谱图重建模糊：高层特征缺乏低层局部纹理信息，导致频谱图边界模糊；多尺度融合可在不同尺度层面增强细节表达。

2) 音源特征差异大：鼓与人声等音源在频谱形态和能量分布上差异明显，单一尺度难以兼顾；多尺度结构可适配多源异质建模。

3) 上采样特征丢失问题：上采样过程存在空间还原不充分的风险，多尺度跳跃连接能有效恢复结构化信息，增强高频细节还原。

3. 实验设计与结果

3.1. 数据集与预处理

实验选用的是音乐源分离标准评测集 MUSDB18 [13]，由 150 首混合音乐样本组成，采样率为 44.1 kHz，音源类别包含 vocals (人声)、drums (鼓)、bass (贝斯)以及其他源(others)。其中，训练集包含 100 首音频，测试集包含 50 首。每首音频都具备完整的混合轨和对应的单独音源轨，为监督训练与精确评估提供了可靠依据。

MUSDB18 数据集覆盖了不同风格、节奏与音色的音乐样本，适用于评估模型在多类音源分离任务中的鲁棒性。

原始音频在输入模型前需进行 STFT (短时傅里叶变换)操作，参数设置为窗长 2048、步长 512。生成的复数谱图被拆解为实部和虚部，并在通道维度进行拼接形成二维输入特征，维度为[B, 2, T, F]。其中 T 为时间帧数，F 为频率 bin (频带数量)。所有样本被分割为长度为 6 秒的片段用于训练。

这种处理方式能有效保留时频信息，使后续模块可以对音源进行更精准的建模与分离。

3.2. 网络参数与模型训练配置

模型使用 PyTorch 框架构建，训练过程中采用 AdamW 优化器，初始学习率设为 1e−4，批大小为 16，训练轮数为 100，见表 1。损失函数采用 SI-SNR + STFT 频谱损失的加权组合，同时考虑时域和频域的误差。为提升训练效率并降低显存占用，使用了半精度混合训练(Mixed Precision)技术。此外，训练中使用早停机制，当验证集 SDR 连续 10 轮无提升时停止训练，防止过拟合。

Table 1. Network parameters and training configuration
表 1. 网络参数与训练配置

名称	型号/参数
操作系统	Ubuntu 20.04 LTS
图像处理器(GPU)	GeForce RTX 3090 (24G)
CPU	Intel (R) Core (TM) i9-12900K
内存	64 G
开发语言	Python 3.9
深度学习框架	PyTorch 1.12.1
CUDA 版本	CUDA 11.6
音频处理库	Librosa 0.9.2, Torchaudio 0.12
初始学习率	1e−4
批大小	16
训练轮数	100

本研究的全部实验在高性能计算服务器上进行，以确保模型训练与评估的稳定性与效率。实验平台采用 Ubuntu 20.04 LTS 操作系统，配置 Intel Core i9-12900K 处理器、64 GB 内存以及 NVIDIA RTX 3090 显卡(24 GB 显存)，具备良好的计算加速能力。软件环境方面，构建于 Python 3.9 基础上，深度学习框架为 PyTorch 1.12.1，配套使用 CUDA 11.6 与 CUDNN 加速库，以实现高效的 GPU 并行计算。音频处理采用 Librosa 0.9.2 与 Torchaudio 0.12，分离质量评估使用 mir_eval、pypesq 与 asteroid.metrics 等常用开源工具包。训练与推理阶段启用了混合精度训练(Mixed Precision Training)与 CUDNN 自动优化策略，以进一步提升资源利用率与训练速度。

整体实验环境配置保证了模型能够在复杂结构下稳定运行，并支持大规模数据训练与高维频谱建模任务的快速迭代。

3.3. 评估指标说明

3.3.1. SDR (Signal-to-Distortion Ratio)

SDR [14]是最常用的音频源分离评估指标之一，用于衡量模型输出信号与目标源信号之间的失真程度。其定义为：$SDR = 10 \cdot \log_{10}\left(\frac{\|s_{target}\|^2}{\|\hat{s} - s_{target}\|^2}\right)\sqrt{a^2+b^2}$，其中 s_{target} 为目标音源信号。SDR 越高表示模型对信号的还原效果越好。本文使用 mir_eval 和 Museval 工具包对 SDR 进行标准计算，适用于不同音源类型(如 vocals、drums 等)。

3.3.2. SI-SNR (Scale-Invariant Signal-to-Noise Ratio)

SI-SNR [15] (Scale-Invariant Signal-to-Noise Ratio，尺度不变信噪比)是语音和音乐源分离任务中常用的客观评估指标之一，旨在度量模型输出信号与参考目标信号之间的相似度，同时消除了对信号幅度(能量)变化的敏感性，从而具有更强的泛化性和稳健性。设真实目标信号为 $s \in \mathbb{R}^T$，模型输出的预测信号为 $\hat{s} \in \mathbb{R}^T$。计算 SI-SNR 的步骤如下：对信号进行去直流(零均值)处理：$s_z = s - \frac{1}{T}\sum_{t=1}^{T} s(t)$，$\hat{s}_z = \hat{s} - \frac{1}{T}\sum_{t=1}^{T} \hat{s}(t)$，其中 s_z 和 $\hat{s}_z$ 分别为去直流后的目标信号与预测信号。将预测信号投影到目标信号方向(构造目标分量)：$s_{\text{target}} = \frac{\langle \hat{s}_z, s_z \rangle}{\|s_z\|^2} \cdot s_z$，该分量表示与目标信号方向一致的部分。计算残差噪声分量：$e_{\text{noise}} = \hat{s}_z - s_{\text{target}}$，该部分表示与目标信号正交的成分，即模型输出中与目标无关的噪声或失真部分。最终计算 SI-SNR 值(单位为 dB)：$\text{SI-SNR} = 10 \cdot \log_{10}\left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2}\right)$，该指标衡量了有用信号分量与噪声分量之间的能量比值，数值越高表示分离质量越好。

3.4. 模型对比实验

为全面评估本文所提出的改进型 TF-GridNet 模型在音乐源分离任务中的性能表现，本文选取了六种当前主流的代表性模型作为对比对象，涵盖了时域、频域及时频联合建模策略，包括 Open-Unmix、Conv-TasNet、TF-GridNet、Demucs v3、SepFormer 和 DiffSep。各模型在 MUSDB18 数据集上分别对 vocals、drums、bass 和 others 四类音源进行分离，计算其 SDR (Signal-to-Distortion Ratio)与 SI-SNR (Scale-Invariant Signal-to-Noise Ratio)两项指标的平均值，并进行对比。

从结果可见(如表 2 所示)，本文提出的改进型 TF-GridNet 模型在所有源类别上的 SDR 表现均优于原始 TF-GridNet，平均 SDR 达到 7.69 dB，相较原始模型提升了 0.58 dB，接近甚至超过 Demucs v3 (7.63 dB)。在人声(vocals)和贝斯(bass)两个音源类别上，本文方法分别达到 8.7 dB 和 7.9 dB，表现尤为突出。

在 SI-SNR 指标方面，本文方法的平均值为 7.93 dB，同样超过了原始 TF-GridNet (7.1 dB)，仅次于 DiffSep (8.18 dB)。考虑到 DiffSep 属于高计算开销的扩散式生成模型，本文方法在保持相对轻量结构的前提下取得了接近甚至更优的性能，展现出良好的性能与效率平衡。

整体来看，改进型 TF-GridNet 在兼顾模型复杂度与分离效果的前提下，在多个典型源类别上均表现出优越的综合性能，验证了所提出结构改进策略的有效性与通用性。

Table 2. SDR comparison results of various audio sources
表 2. 各音源 SDR 对比结果

模型	类型	Vocal	Drums	Bass	Others	平均 SDR
Open-Unmix [3]	STFT-based	8.0	6.7	7.1	6.0	6.95
Conv-TasNet [5]	Time-based	7.9	6.4	6.7	5.9	6.73
TF-GridNet (原始) [9]	TF-dual	8.1	6.9	7.3	6.1	7.1
Demucs v3 [4]	Hybird	8.4	7.2	7.8	7.1	7.63
SepFormer [7]	Transformer	8.6	7.4	7.9	7.3	7.8
DiffSep [8]	DIffusion	8.9	7.8	8.4	7.6	8.18
改进型 TF-GridNet	TF-dual + 增强	8.7	7.4	7.9	6.7	7.68

Table 3. SI-SNR comparison results of various audio sources
表 3. 各音源 SI-SNR 对比结果

模型	类型	Vocal	Drums	Bass	Others	平均 SI-SNR
Open-Unmix [3]	STFT-based	6.2	5.7	5.1	4.8	5.45
Conv-TasNet [5]	Time-based	7.1	6.6	6.4	5.9	6.5
TF-GridNet (原始) [9]	TF-dual	8.1	6.9	7.3	6.1	7.1
Demucs v3 [4]	Hybird	8.7	7.4	7.6	7.6	7.83
SepFormer [7]	Transformer	9.0	7.8	7.6	7.1	7.88
DiffSep [8]	DIffusion	9.3	8.1	7.9	7.4	8.18
改进型 TF-GridNet	TF-dual + 增强	8.9	7.8	7.9	7.1	7.93

此外，本文进一步对各类音源的分离性能进行了单独分析。从表 2 和表 3 中可见，改进型 TF-GridNet 在不同音源类别上均表现出良好的适应性与建模能力，特别在人声(vocals)与贝斯(bass)两个类别上提升显著。人声分离任务中，本文模型在 SDR 和 SI-SNR 两项指标上分别达到 8.7 dB 和 8.9 dB，较原始 TF-GridNet 分别提升 0.6 dB 和 0.8 dB，超越 Demucs v3 与 SepFormer，表现出对主旋律信号的更强捕捉能力。贝斯作为低频主导音源，本文模型通过多尺度解码器增强了低频建模的层次表达，SDR 达到 7.9 dB，在对比模型中位居前列，仅次于 DiffSep。drums 类音源中，模型得益于通道注意力机制对节奏敏感通道的强化，在保持较小模型复杂度的前提下，分离质量与高阶 Transformer 结构模型相当。others 类音源中，由于包含大量伴奏与环境声，通常分离难度较高，但改进型模型仍实现 6.7 dB SDR 和 7.1 dB SI-SNR，优于 TF-GridNet 与 Conv-TasNet，体现出较强的泛化性与抗干扰能力。上述结果从不同音源维度进一步验证了本文所提结构优化策略对提升分离性能的有效性和稳定性。

3.5. 消融实验

为进一步验证本文提出的结构性改进在模型性能中的独立贡献，设计了模块级别的消融实验。实验

分别在 TF-GridNet 的基础上，逐步引入 ECA 通道注意力模块、门控残差融合结构以及多尺度解码器，并评估其在 SDR 与 SI-SNR 两项指标下的分离性能表现。

Table 4. SDR ablation study
表 4. SDR 消融实验

模型	ECA	门控残差	多尺度解码器	Vocal	Drums	Bass	Others	平均 SDR
TF-Gridnet				8.1	6.9	7.3	6.1	7.1
+ECA	√			8.5	7.1	7.6	6.3	7.38
+残差门控结构		√		8.4	7.0	7.5	6.3	7.3
+多尺度解码器			√	8.6	7.3	7.8	6.5	7.55
Ours	√	√	√	8.7	7.4	7.9	6.7	7.68

Table 5. SI-SNR ablation study
表 5. SI-SNR 消融实验

模型	ECA	门控残差	多尺度解码器	Vocal	Drums	Bass	Others	平均 SI-SNR
TF-Gridnet				8.7	7.4	7.9	6.5	7.63
+ECA	√			9.0	7.6	8.2	6.7	7.88
+残差门控结构		√		8.9	7.5	8.1	6.7	7.8
+多尺度解码器			√	9.1	7.8	8.4	6.9	8.05
Ours	√	√	√	9.2	7.9	8.6	7.1	8.2

从表 4 和表 5 中可以看出，引入 ECA 模块后，模型的平均 SDR 从 7.10 dB 提升至 7.38 dB，SI-SNR 也由 7.63 dB 提升至 7.88 dB，说明通道注意力机制能够有效增强网络对关键信道的选择性，特别是在高频信息显著的人声分离任务中表现突出。门控残差结构的引入也带来了正向提升，平均 SDR 增加 0.2 dB，SI-SNR 增加 0.1 7dB，验证了其在控制主路径与残差路径信息融合方面的灵活性与建模能力。此外，多尺度解码器的加入显著提升了高频还原能力，平均 SDR 达到 7.55 dB，SI-SNR 提升至 8.05 dB，表明多尺度路径与跳跃连接能够弥补上采样阶段的结构性信息丢失。在三项改进模块全部引入的情况下，模型达到最佳性能，平均 SDR 为 7.68 dB，SI-SNR 达到 8.2 dB，验证了多模块协同设计的合理性与有效性。

整体而言，本文提出的每项结构改进均在不同维度提升了模型的音频建模与重建能力，且具有良好的叠加互补性，为最终模型性能的全面提升提供了支撑。

4. 结论

本文在 TF-GridNet 基础上提出了融合轻量通道注意力、门控残差连接与多尺度解码结构的改进型音乐源分离模型。实验在 MUSDB18 数据集上开展，并与多种主流方法进行了对比分析。结果表明，改进型 TF-GridNet 在多个音源类别上取得了更优的分离性能，特别是在 Bass 和 Others 这类细节依赖度较高的音源上表现出显著提升。同时，通过引入通道选择机制与多尺度重建路径，模型在保持感知质量的同时，还具备更强的特征建模能力与重建能力。此外，消融实验验证了三项结构性优化的有效性，每一项改进均对模型性能产生了正向影响，且组合使用时可实现整体性能的最优化。

参考文献

[1] Comon, P. (1994) Independent Component Analysis, a New Concept? *Signal Processing*, **36**, 287-314.

https://doi.org/10.1016/0165-1684(94)90029-9

[2] Virtanen, T. (2007) Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech and Language Processing*, **15**, 1066-1074. https://doi.org/10.1109/tasl.2006.885253

[3] Stöter, F., Uhlich, S., Liutkus, A. and Mitsufuji, Y. (2019) Open-Unmix—A Reference Implementation for Music Source Separation. *Journal of Open Source Software*, **4**, Article No. 1667. https://doi.org/10.21105/joss.01667

[4] Défossez, A., Usunier, N., Bottou, L. and Bach, F. (2021) Hybrid Spectrogram and Waveform Source Separation. *Advances in Neural Information Processing Systems* 34: *Annual Conference on Neural Information Processing Systems* 2021, *NeurIPS* 2021, 6-14 December 2021.

[5] Luo, Y. and Mesgarani, N. (2019) Conv-Tasnet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**, 1256-1266. https://doi.org/10.1109/taslp.2019.2915167

[6] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M. and Zhong, J. (2021) Attention Is All You Need in Speech Separation. 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Toronto, 6-11 June 2021, 21-25.

[7] ITU-T (1996) Recommendation P.800: Methods for Subjective Determination of Transmission Quality. International Tele-Communication Union.

[8] Luo, Y., Chen, J., Du, J. and Yoshioka, T. (2023) DiffSep: Leveraging Diffusion Models for Speech Separation. *IEEE Proceedings of ICASSP* 2023, Rhodes Island, 4-10 June 2023, 1-5.

[9] Luo, Y., Lin, Z.-Q., Zhang, J. and Mesgarani, N. (2022) TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation. *Proceedings of Interspeech* 2022, Incheon, 18-22 September 2022, 2768-2772.

[10] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 11531-11539. https://doi.org/10.1109/cvpr42600.2020.01155

[11] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[12] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., *et al*., Eds., *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015, Springer International Publishing, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[13] Rafii, Z., Liutkus, A., Stoeter, F.-R., Mimilakis, S.I. and Bittner, R. (2017) The MUSDB18 Corpus for Music Separation. Machine Learning for Signal Processing (MLSP). https://sigsep.github.io/datasets/musdb.html

[14] Vincent, E., Gribonval, R. and Fevotte, C. (2006) Performance Measurement in Blind Audio Source Separation. *IEEE Transactions on Audio, Speech and Language Processing*, **14**, 1462-1469. https://doi.org/10.1109/tsa.2005.858005

[15] Le Roux, J., Weiss, R.J. and Kinoshita, K. (2019) SNR-Based Objective Evaluation of Source Separation Methods. *IEEE Transactions on Audio, Speech, and Language Processing*, **27**, 929-941.