http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0019

特征表达能力增强的声音事件定位与检测网络

章东平[1,*]，符珍涛[1]，王杼涛[1]，林丽莉[2]，魏明[3]

(1. 中国计量大学 信息工程学院，杭州 310018；2. 浙江工商大学 信息与电子工程学院，杭州 310018；
3. 杭州爱华智能科技有限公司，杭州 311121)

摘 要：针对传统深度学习模型难以捕捉输入特征图中的长上下文特征关联及通道与空间维度上的关键特征信息，导致声音事件定位与检测 (SELD) 错误率高、性能不理想的问题，基于声学场景分类和声音事件检测挑战赛中的基线模型 SELDnet，提出一种基于增强特征表达能力的声音事件定位与检测网络 (FE-SELDnet)。采用组归一化和 SiLU 激活函数来解决函数无法反向传播导致神经元死亡的问题；引入卷积块注意力模块 (CBAM) 来捕捉声学特征中通道与空间 2 个维度的重要特征，抑制不必要的特征，加强网络对特征信息的敏感性和准确性，提高信息流动；引入 Transformer 模块来捕获更长的语音上下文特征关联，并结合局部特征，提升模型在声音事件定位与检测任务中的精确性和鲁棒性。在 TUT Sound Events 数据集上的实验结果表明：FE-SELDnet 与基线网络性能相比有较大的提升，错误率从 0.45 降低到 0.326，SED 评分和 DOA 评分分别从 0.45 和 0.32 降至 0.26 和 0.25，F_1 分数提高到 79.4%，验证了 FE-SELDnet 具有更高的优越性。

关 键 词：声音事件定位与检测；特征表达增强；注意力机制；深度学习；组归一化
中图分类号：TN912.3; TP181
文献标志码：A **文章编号**：1001-5965(2026)04-1088-08

声音事件定位与检测(sound event localization and detection, SELD)可以分为声音事件检测(sound event detection, SED)和声源定位 2 个独立任务，旨在检测声音事件及其在时间上的活动情况，并在活动时估计其空间位置。SELD 在很多应用中发挥了重要作用，如智能家居、生物多样性监测、智能汽车及异常声音事件检测等领域，是当前人工智能领域的研究热点[1-2]。

使用模板匹配法[3]结合隐马尔可夫模型、高斯混合模型等传统模型进行声音事件检测时很难处理声音混叠的问题。为解决重叠事件识别问题，Heittola、Mesaros 等利用多个受限 Viterbi 通路检测多个重叠的声音事件[4]。基于非负矩阵分解原理的方法也能处理混叠的声音事件[5]，但不能处理声音事件时域的相关性信息，导致检测效果不理想。为提升 SED 性能，支持向量机[6]被用于对声音事件进行检测，虽然能够提高检测准确率，但训练过程烦琐，训练时间长。

传统的声源定位方法有到达时间差、高分辨率谱估计、可控波束形成及声强估计等[7-9]，其在噪声环境下的性能受限，表现为特征表达力欠佳、精度不高、误差较大等问题。

深度学习网络结构，如卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)[10]，被用于环境声音识别领域，取得了相对较高的识别准确率。CNN 能够进行局部特征的学习，而 RNN 能够对时间上下文特征联系进行学习，于是，结合 CNN 与

收稿日期：2024-01-11；录用日期：2024-02-29；网络出版时间：2024-03-15 15:16
网络出版地址：link.cnki.net/urlid/11.2625.V.20240314.2102.001
基金项目：浙江省重点研发计划 (2023C01034,2023C01030,2023C01032)
*通信作者. E-mail：06a0303103@cjlu.edu.cn

引用格式：章东平，符珍涛，王杼涛，等. 特征表达能力增强的声音事件定位与检测网络 [J]. 北京航空航天大学学报，2026，52（4）：1088-1095. ZHANG D P，FU Z T，WANG Z T，et al. Sound event localization and detection network with enhanced feature expression [J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（4）：1088-1095 （in Chinese）.

RNN 产生的 CNN-LSTM[11] 网络同时具备 2 种网络的学习能力。文献 [12-13] 设计的 RD3Net 网络结构采用了多层 CNN 后接高效门控循环单元(gated recurrent unit, GRU)的方式, 以提升模型性能。鉴于传统的 RNN 网络普遍存在训练耗时较长、参数规模较大及训练难度高等问题, 文献 [14] 提出了基于 Transformer 的 SELD 模型, 显著提高了模型训练与推理效率。也有研究尝试改进 CNN 使其产生 RNN 的学习功能。例如, Bai 等[15] 提出了时序卷积网络 (temporal convolutional network, TCN), 网络结构含有学习时间上下文特征信息的能力。文献 [16] 提出了 SELDEnet 模型方案, 能够处理多通道音频输入, 通过 CNN 提取局部信息, 借助 GRU 学习语音信号时间上的信息。SELDEnet 模型被用于声音领域权威比赛 DCASE 挑战赛中任务 3 的基线模型。

考虑到在 SELD 中, 声学特征的提取对模型预测能力和最终分类结果具有重要影响, 本文基于 SELDnet 基线模型, 提出了一种基于增强特征表达能力的 SELD 网络 (feature enhanced SELD network, FE-SELDnet)。采用组归一化(group normalization, GN)结合 SiLU 激活函数, 同时引入卷积块注意力模块(convolutional block attention module, CBAM)及 Transformer 模块, 通过 3 个方面的改进, 增强网络对语音特征的提取, 提升了模型在 SELD 任务中的精确性和鲁棒性。

1　相关工作

近年来, 深度学习技术在计算机视觉、自然语言处理、语音与音频处理等领域中应用效果突出, 对 SELD 性能的提升起到了积极推动作用。研究初期, 深度神经网络[17] 成为涉足这一领域的首批尝试。随后, CNN[18-19] 与 RNN[20-21] 也相继被引入 SELD 任务。2015 年, 由 Hirvonen 团队[22] 开展的研究表明, 尽管 CNN 可用于声音事件定位与检测, 但在面对多事件重叠场景时, 其预测准确性不尽如人意, 误差显著增大。原因在于: CNN 善于提取局部特征, 而对于时序及语义信息的学习能力相对较弱。鉴于此, 研究者们结合 CNN 与 RNN 的优势, 构建了卷积循环神经网络 (convolutional recurrent neural network, CRNN)[23] 模型, 并取得了明显的性能提升。在此基础上, 为进一步提升网络获取音频信息的时域、频域及通道间特征能力, 研究者们对 CRNN 模型进行了扩展, 采用三维卷积操作对输入的多通道数据进行联合分析, 从而提高了对复杂声学场景下 SELD 的效能[24]。

由坦佩雷理工大学主办的 DCASE 挑战赛作为一个国际公认的声学研究平台, 汇聚了全球众多学者的关注与积极参与。自 2019 年起, 该挑战赛中的任务 3 着重聚焦于 SELD 问题, 吸引了大量研究者投入, 并由此催生出一系列创新的方法和模型。例如, 曹寅等[25] 提出了一种双阶段策略, 通过训练独立的 SED 和方向到达角(direction of arrival, DOA)估计模型, 将 SED 模型的输出作为掩模指导 DOA 模型预测, 该方法在提升 SELD 性能方面有所突破, 并在赛事中取得了良好的成绩。Ranjan 等[26] 研发了一种基于残差网络架构的 RNN 算法, 有效提高了 SELD 性能, 并防止了网络退化。此外, Nguyen 等[27] 提出了一个基于迁移学习融合 RNN 网络的通用 SELD 框架, 先通过预训练分别解决 SED 和 DOA 估计, 再利用循环层将 2 部分的估计结果有机融合以生成最终的 SELD 输出。Zhang 等[28] 设计了一种结合 CNN 与 Conformer 模块的 CNN-Conformer 模型, 该模型能够捕捉并利用时间上下文特征信息, 成功完成了 SELD 任务。Lee 等[29] 提出了 CMA-SELD 模型, 该模型利用跨模态注意力 (cross-modal attention, CMA) 机制来学习和关联 SED 与声源定位特征, 并采用参数共享策略提取用于 SELD 的必要特征, 从而提升了 SED 和声源定位的整体性能。

2　FE-SELDnet 网络

本文提出的 FE-SELDnet 模型的整体架构布局如图 1 所示。

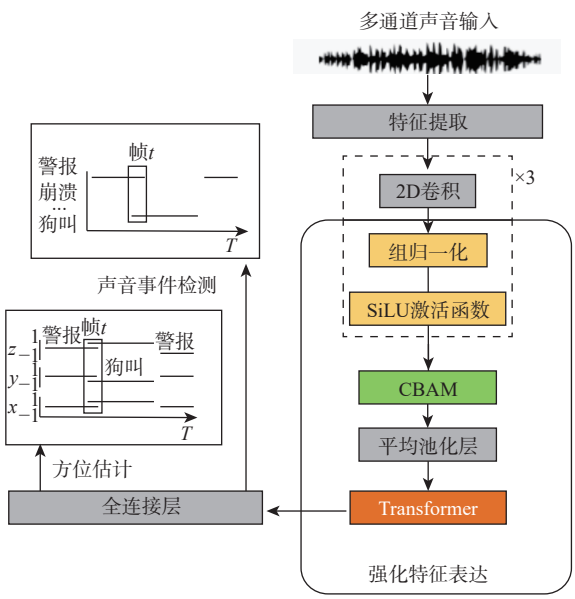

图 1　本文网络的总体结构

Fig. 1　Overall structure of the proposed network

FE-SELDnet 网络主要组成部分如下: 首先, CNN 模块通过采用组归一化和 SiLU 激活函数协同

卷积操作，以提取声音信号的局部特征信息。其次，加入CBAM模块，旨在强化对声学特征中重要信息的表达能力。然后，引入Transformer模块，能够有效捕获并整合更长时段的上下文特征关联。最后，通过上述3个关键模块的有机组合与协作，FE-SELDnet模型在特征提取与表达层面得到了显著增强。这些经过深度处理和转化的特征信息将通过全连接层进行整合与分类，以实现声音事件的精准检测与定位。

2.1 SELDnet基线模型

SELDnet是一种专为基于麦克风阵列输入的复杂声场环境设计的SELD一体化神经网络模型，其架构如图2所示，主要由2部分构成：CNN和双向RNN。CNN部分，网络架构包括3个连续的卷积层块及1个平均池化层。每个卷积层块内部集成了多个核心计算单元，即使用3×3大小的卷积核进行特征提取，辅以批归一化（batch normalization, BN）技术以稳定训练过程，并采用ReLU激活函数引入非线性特性。双向RNN部分，模型使用BiGRU这一RNN变体结构，该结构能够同步考虑序列数据的过去与未来信息，从而在预测过程中充分利用时间维度上的上下文信息，一定程度上提升了模型在SELD任务上的性能表现。

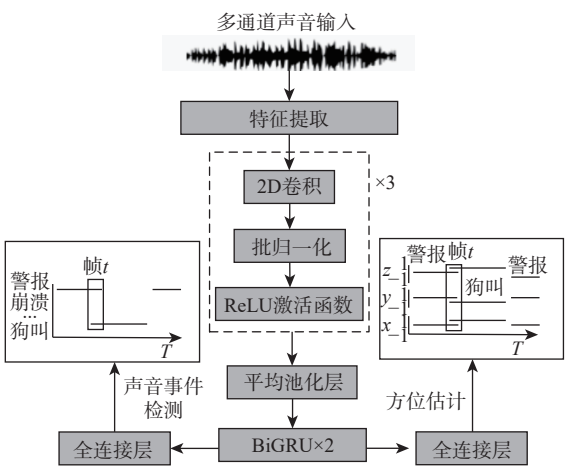

图2 SELDnet网络结构

Fig. 2 SELDnet network structure

2.2 组归一化与SiLU激活函数

在网络训练进程中，模型参数经历迭代更新，其中，当前层参数的变化受到前一层参数变动的影响，因此，可能导致浅层神经网络在反向传播中出现梯度消失问题，从而影响网络的收敛性能。为缓解梯度消失难题，网络结构通常采用批归一化技术，将输入数据转化为近似标准正态分布的形式，增强非线性函数对输入数据的响应灵敏度，进而加速网络收敛速度并提高预测准确性。但是，批归一化会因批量大小而影响效果，尤其是在小批量情况下，批归一化的应用可能导致输出结果的误差增大。因此，合理选择批量大小对保证批归一化效果的准确性至关重要。相比之下，组归一化[30]对通道方向归一化，将特征图按预先设定的组数进行分割，每组独立计算各自的均值和方差。与批量大小无关的特性使得组归一化能够确保模型在不同批量大小下仍能够有效加快网络收敛速度，并维持稳定的性能表现。本文模型利用组归一化特性，以分组形式对特征图进行归一化处理，从而提升模型训练效率与准确性。

ReLU激活函数将所有负值都设为0，正向梯度设为1，大幅提高网络模型的计算速度和收敛速率，但也使得输入值为负值时，函数负方向的梯度为0，反向传递失败，部分神经元“死亡”，削弱了网络的学习能力。为解决上述问题，本文采用SiLU激活函数，其是Swish[31]激活函数的特例，表示为

$$f(x) = x \cdot \mathrm{sigmoid}(x) \tag{1}$$

如图3所示，曲线平滑，无上界有下界，且非单调，输入信息小于0时也不会丢失，能够加强特征信息的表达，对SELD网络性能可以发挥有利作用，对网络的预测效果相较ReLU函数更好，特别是在深层网络中优势更加明显。

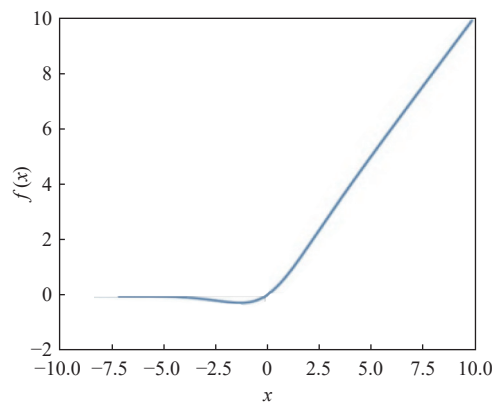

图3 SiLU激活函数

Fig. 3 SiLU activation function

2.3 CBAM模型

在SELD任务中，语音特征不仅涵盖通道层面的特性，还蕴含着时间序列上的动态演变信息。在丰富的信息集合中，区分有效信息与冗余信息至关重要，目的是使有限的计算资源优先服务于有价值的信息提取。通道注意力与时空注意力机制恰好能从输入特征图的通道维度和时空维度甄别并聚焦重要特征。因此，本文提出通道时空注意力模块[32]，通过整合通道注意力子模块与时空注意力子模块，对特征图的通道维度和时空维度进行注意力推测，

并将由此生成的注意力图与原始特征图相乘, 以实现对特征的自适应优化。CBAM 作为轻量级且通用性强的模块, 可以便捷地嵌入到现有的神经网络结构中, 从而为 SELD 等任务提供精细化的注意力导向与特征提取能力。CBAM 模块结构如图 4 所示。

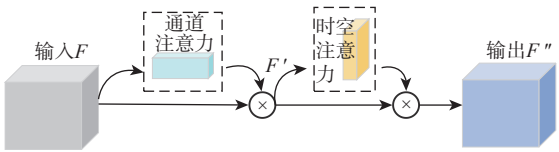

图 4　CBAM 模块结构

Fig. 4　CBAM structure

2.4　Transformer 网络结构

Transformer 网络模型主要由编码器和解码器组成。其中, 解码器[33]模块如图 5 所示, 核心部分是多头自注意力模块和前馈神经网络。

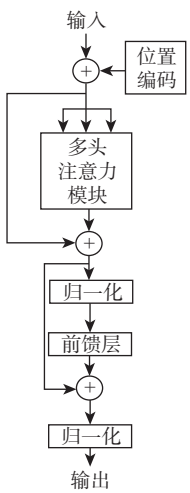

图 5　Transformer 解码器结构

Fig. 5　Transformer decoder structure

多头自注意力模块的实现如下:

$$\begin{cases} \boldsymbol{Q}=\boldsymbol{W}^{Q}\boldsymbol{I} \\ \boldsymbol{K}=\boldsymbol{W}^{K}\boldsymbol{I} \\ \boldsymbol{V}=\boldsymbol{W}^{V}\boldsymbol{I} \end{cases} \tag{2}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\boldsymbol{V}\right) \tag{3}$$

式中: $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别为查询向量、键向量和值向量, 由网络接收的输入向量 $\boldsymbol{I}$ 分别通过与对应的权值矩阵 $\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$和 $\boldsymbol{W}^{V}$进行矩阵乘法运算得出; d_k 为缩放因子, 用于计算 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的点积后进行归一化处理, 以防止内积结果的数值过大或过小影响注意力分配的稳定性。

鉴于 SELD 任务不需要依据前一时刻的声音事件去预测后一时刻的声音事件, 本文仅选用 Transformer 结构中的解码器部分进行构建和优化。

3　实验结果与分析

3.1　数据集与评价指标

本文采用 TUT Sound Events 数据集中的混响和合成冲激响应数据集 RESYM。

采用错误率、F_1 分数、SED 评分、DOA 评分这 4 个评价指标来评价 SELD 的预测性能。

$$F_1=\frac{2\sum_{k=1}^{K}T_{\mathrm{P}}(k)}{2\sum_{k=1}^{K}T_{\mathrm{P}}(k)+2\sum_{k=1}^{K}F_{\mathrm{P}}(k)+2\sum_{k=1}^{K}F_{\mathrm{N}}(k)} \tag{4}$$

式中: K 为每秒的帧总数; T_{P} 表示每秒正确识别为正类的帧数; F_{P} 表示每秒误判为正类的帧数; F_{N} 表示每秒未能正确识别为正类的帧数。

$$R_{\mathrm{E}}=\frac{\sum_{k=1}^{K}S(k)+\sum_{k=1}^{K}D(k)+\sum_{k=1}^{K}G(k)}{\sum_{k=1}^{K}N(k)} \tag{5}$$

式中: R_{E} 为错误率; $S(k)$为第 k 时刻所需的帧替换数量; $D(k)$为第 k 时刻待删除的帧数; $G(k)$为需要在第 k 时刻补充的帧数目; $N(k)$为每秒内的总帧数。

$$S(k)=\min(F_{\mathrm{N}}(k),F_{\mathrm{P}}(k)) \tag{6}$$

$$D(k)=\max(0,F_{\mathrm{N}}(k)-F_{\mathrm{P}}(k)) \tag{7}$$

$$G(k)=\max(0,F_{\mathrm{P}}(k)-F_{\mathrm{N}}(k)) \tag{8}$$

在声音事件定位中, 将实际三维笛卡儿坐标定义为 $(x_{\mathrm{R}}, y_{\mathrm{R}}, z_{\mathrm{R}})$, 而预测出的对应三维笛卡儿坐标则记为 $(x_{\mathrm{P}}, y_{\mathrm{P}}, z_{\mathrm{P}})$。对于每个坐标轴, 可以通过计算预测误差 $\Delta x=x_{\mathrm{R}}-x_{\mathrm{P}}$、$\Delta y=y_{\mathrm{R}}-y_{\mathrm{P}}$、$\Delta z=z_{\mathrm{R}}-z_{\mathrm{P}}$ 来量化预测位置与真实位置间的偏差程度, 将偏差程度定义为 E_{L}, 其计算公式为

$$E_{\mathrm{L}}=\frac{1}{M}\sum_{m=1}^{M}2\arcsin\left(\frac{\sqrt{\Delta x^2+\Delta y^2+\Delta z^2}}{2}\right)\frac{180}{\pi} \tag{9}$$

式中: M 为整个时间段内包含声音事件的有效帧总数。

SED 评分 S_{s} 与 DOA 评分 D_{s} 是 SELD 任务的整体性能指标, 其计算公式为

$$S_{\mathrm{s}}=[R_{\mathrm{E}}+(1-F_1)]/2 \tag{10}$$

$$D_{\mathrm{s}}=[E_{\mathrm{L}}/180+(1-R_{\mathrm{F}})]/2 \tag{11}$$

式中: R_{F} 为帧召回率。

在理想状态下, R_{E}、SED 评分、DOA 评分越低, F_1 分数越高, SELD 网络的性能越好。

3.2 实验结果与对比

实验依托 Linux 操作系统环境，在 Ubuntu 20.04 版本上运行，采用 Python 3.8 编程语言和 TensorFlow 库的 2.5.1 版本作为开发工具和技术框架，在 NVIDIA Tesla V100，显存为 32 GB 的 GPU 上实现。网络学习率设置为 0.000 1,迭代轮数为 120 轮，批大小为 32。

为严谨验证本文提出的改进措施对模型性能的实际提升效果，通过实验，根据评价指标逐一分析并评估改进措施对模型在 SELD 任务上的具体贡献。

如图 6 和图 7 所示，在训练过程中，观察到训练损失和验证损失逐步递减并趋向平稳，显示了网络成功地从输入数据中学习到了关键特征，进而有效实现了声音事件识别任务。同时，F_1 分数上升并最终进入收敛阶段，进一步证实了方法性能不断优化，最终达到了稳定状态。

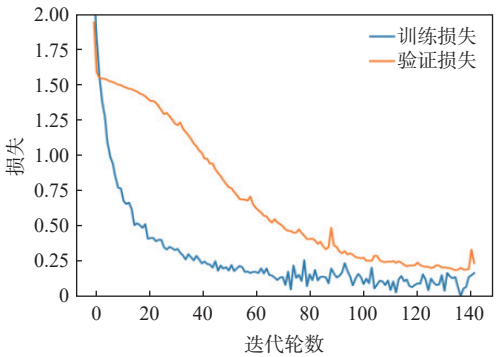

图 6 训练与验证损失

Fig. 6 Training and validation loss

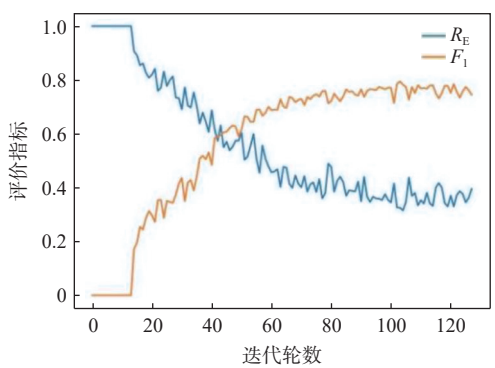

图 7 评价指标随迭代轮数变化

Fig. 7 Evaluation metrics changes with iteration rounds

图 8 显示，本文方法的系统总体评分 (SELD)、位置识别系统评分 (DOA)、事件识别系统评分 (SED) 随着迭代轮数逐渐增大都逐步趋于稳定值。

为分析 FE-SELDnet 网络性能，选择在 SELD 研究中的若干神经网络模型进行实验对比，其中，包括 CRNNnet[34]、CNN-Conformer 及 M2MAST[35] 3 种网络模型。通过实验对比，利用各模型实验得出评价指标，对 FE-SELDnet 的性能进行了深入的分析和验证。

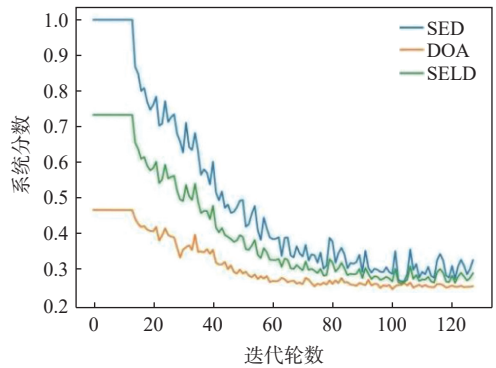

图 8 系统分数随迭代轮数变化

Fig. 8 System score changes with iteration rounds

1）CRNNnet: 采用 CRNN 作为主体架构,通过去除检测分支简化了整体架构，提高了性能。

2）CNN-Conformer: 利用 Conformer 模块替换 SELDnet 基线网络中的 BiGRU 模块，提出通过联合 CNN 与 Transformer 对全局和局部上下文依赖性建模。

3）M2MAST: 使用迁移学习方法，其权重通过修改预训练权重获得，提出多对多音频频谱变换器 (audio spectrogram transformer, AST) 以满足 SELD 的多通道音频输入与结果输出。

表 1 为 3 种模型与本文模型在 TUT Sound Events 数据集上的实验结果，分别对比了 F_1 分数、错误率、DOA 评分、SED 评分指标。可以看出，本文模型错误率较其他模型分别减少 0.102、0.070、0.048；F_1 分数分别增加了 8.2%、7%、5.3%；DOA 评分和 SED 评分这 2 项综合评分与其他模型比较分别降低了 0.17、0.14、0.125 和 0.05、0.041、0.031。实验结果进一步证实了本文方法可以增强特征表达能力，实现高性能的声音事件检测与定位。

表 1 不同模型的评价指标

Table 1 Evaluation indexes of different models

方法	错误率↓	F_1分数/%↑	DOA 评分↓	SED 评分↓
CRNNnet	0.428	71.2	0.42	0.31
CNN-Conformer	0396	72.4	0.39	0.301
M2MAST	0.374	74.1	0.375	0.291
FE-SELDnet（本文）	**0.326**	**79.4**	**0.25**	**0.26**

注：R_E、SED评分、DOA评分越低，F_1分数越高，SELD网络的性能越好，数据加黑表示性能最优。

3.3 消融实验

为验证每个改进部分的有效性,分别验证每个部分对模型性能的影响。表 2 分别对比了 GN、SiLU、CBAM、Transformer 模块对模型的综合性能的影响。可以看出，在加入 CGS 模块后，错误率、F_1 分数、DOA 评分、SED 评分指标均有明显提

表 2　消融实验结果比较

Table 2　Results comparison of ablation experiment

模型	错误率↓	F_1 分数/%↑	DOA 评分↓	SED 评分↓
SELDnet	0.45	68.7	0.32	0.45
SELDnet+GN、SiLU	0.34	76.6	0.27	0.34
SELDnet+GN、SiLU+CBAM	0.325	78.5	0.26	0.32
SELDnet+GN、SiLU+CBAM+Transformer	0.326	79.4	0.25	0.26

升。在此基础上加入 CBAM 结构，错误率下降了 0.015, F_1 分数上升了 1.9%, DOA 评分、SED 评分分别下降了 0.01、0.02。而在加入 Transformer 结构后，错误率稍有上升，其他指标均有小幅度提升。实验结果表明，对于 SELD，提出的 3 部分特征表达增强的方法都是有效的：①采用组归一化和 SiLU 激活函数来解决函数无法反向传播导致神经元死亡的问题，提高局部特征表达。②引入 CBAM 模块捕捉声学重要特征，抑制无关特征，加强网络对重要特征信息表达，提高信息流动。③引入 Transformer 模块来捕获更长的语音上下文特征关联，并结合局部特征，学习到了更多的语境信息，提升模型在 SELD 任务中的精确性和鲁棒性。通过 3 个部分的改进，增强了特征表达能力，提高了 SELD 性能。

4　结　论

1）引入组归一化与 SiLU 激活函数相结合的特征提取系统增强特征提取，与基线网络 SELDnet 相比，各项评价指标都有所提升，尤其 F_1 分数从 68.7% 提升到 76.6%。

2）引入 CBAM 模块来捕捉通道与时空维度的重要声学特征，实验结果得出，错误率降低到 0.325，比只引入组归一化与 SiLU 激活函数时降低了 0.015，比基线网络降低了 0.125。

3）引入 Transformer 模块来捕获更长的语音上下文特征关联，并结合局部特征，提升精确性。相较于基线网络模型，F_1 分数从 68.7% 上升到 79.4%，上升了 10.7%，DOA 评分从 0.32 下降到 0.25，下降了 0.07，SED 评分从 0.45 下降到 0.26，下降了 0.19。

在 SELD 任务中，输入深度学习网络的语音特征是经过处理后的特征，并不包含原始语音信号的所有特征信息，还有大量的信息不包括，因此也会对后续的 SELD 结果产生一些负面影响。在下一步的研究中，准备向输入特征方向进行研究，探索多种语音特征，进行融合等处理步骤，以获得包含更全面的语音信息，增强网络的性能。

参考文献（References）

[1] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[2] DABRAN I, ELMAKIAS O, SHMELKIN R, et al. An intelligent sound alam recognition system for smart cars and smart homes [C]//Proceedings of the IEEE/IFIP Network Operations and Management Symposium. Piscataway: IEEE Press, 2018: 1-4.

[3] SCHRÖDER J, MORITZ N, SCHÄDLER M R, et al. On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events' [C]//Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Piscataway: IEEE Press, 2013: 1-4.

[4] HEITTOLA T, MESAROS A, ERONEN A, et al. Context-dependent sound event detection[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2013, 2013: 1.

[5] KOMATSU T, TOIZUMI T, KONDO R, et al. Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop. Budapest: [s.n.], 2016: 45-49.

[6] XU X Y, YU J D, CHEN Y Y, et al. Leveraging audio signals for early recognition of inattentive driving with smartphones[J]. IEEE Transactions on Mobile Computing, 2018, 17(7): 1553-1567.

[7] VELÁZQUEZ I M, REN Y, HANEDA Y, et al. A fusion method based on class rotations for DNN-DoA estimation on spherical microphone array[C]//Proceedings of the 29th European Signal Processing Conference. Piscataway: IEEE Press, 2021: 885-889.

[8] 鄢社锋, 马远良, 侯朝焕. 宽带波束域相干信号子空间高分辨方位估计[J]. 声学学报, 2006, 31(5): 418-424.
YAN S F, MA Y L, HOU C H. High resolution azimuth estimation of coherent signal subspace in broadband beam domain[J]. Journal of Acoustics, 2006, 31(5): 418-424(in Chinese).

[9] 李伟红, 汤海兵, 龚卫国. 公共场所异常声源定位中时延估计方法研究[J]. 仪器仪表学报, 2012, 33(4): 750-756.
LI W H, TANG H B, GONG W G. Research on time delay estimation method for abnormal sound source location in public places[J]. Chinese Journal of Instrumentation, 2012, 33(4): 750-756(in Chinese).

[10] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2014: 1724-1734.

[11] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling [EB/OL](2018-04-19)[2024-01-01]https://arxiv.org/abs/1803.01271.

[12] SHIMADA K, TAKAHASHI N, TAKAHASHI S, et al. Sound event localization and detection using activity-coupled cartesian DOA vector and RD3Net[EB/OL]. (2020-07-31)[2024-01-01]. https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Shimada_139.pdf.

[13] TAKAHASHI N, MITSUFUJI Y. D3Net: densely connected multi-dilated DenseNet for music source separation[EB/OL]. (2021-05-27)

[2024-01-01]. https://arxiv.org/abs/2010.01733v4.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2024-01-01]. https://arxiv.org/abs/1706.03762.

[15] BAI S J, KOLTER J Z, KOLTUN V. Trellis networks for sequence modeling[EB/OL]. (2019-05-11)[2024-01-01]. https://arxiv.org/abs/1810.06682.

[16] ADAVANNE S, POLITIS A, NIKUNEN J, et al. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(1): 34-48.

[17] SAMEK W, BINDER A, MONTAVON G, et al. Evaluating the visualization of what a deep neural network has learned[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(11): 2660-2673.

[18] BATTAGLINO D, LEPAULOUX L, EVANS N. Acoustic scene classification using convolutional neural networks[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events. Piscataway: IEEE Press, 2016: 1-5.

[19] ZINEMANAS P, CANCELA P, ROCAMORA M. End-to-end convolutional neural networks for sound event detection in urban environments[C]//Proceedings of the 24th Conference of Open Innovations Association. Piscataway: IEEE Press, 2019: 533-539.

[20] HAYASHI T, WATANABE S, TODA T, et al. Duration-controlled LSTM for polyphonic sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(11): 2059-2070.

[21] ZÖHRER M, PERNKOPF F. Virtual adversarial training and data augmentation for acoustic event detection with gated recurrent neural networks[C]//Interspeech 2017. [S.l.]: ISCA, 2017: 493-497.

[22] HIRVONEN T. Classification of spatial audio location and content using convolutional neural networks[C]//Audio Engineering Society Convention 138. [S.l.]: Audio Engineering Society, 2015: 1-10.

[23] GRUMIAUX P A, KITIĆ S, GIRIN L, et al. A survey of sound source localization with deep learning methods[J]. Journal of the Acoustical Society of America, 2022, 152(1): 107-151.

[24] MEI P C, YANG J B, ZHANG Q, et al. A method of sound event localization and detection based on three-dimension convolution[C]//Proceedings of the 7th International Conference on Image, Vision and Computing. Piscataway: IEEE Press, 2022: 872-878.

[25] CAO Y, KONG Q, IQBAL T, et al. Polyphonic sound event detection and localization using a two-stage strategy[EB/OL]. (2019-11-05)[2024-01-01]. https://arxiv.org/abs/1905.00268.

[26] RANJAN R, JAYABALAN S, NGUYEN T N T, et al. Sound event detection and direction of arrival estimation using ResidualNet and recurrent neural networks[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop. [S.l.]: DCASE, 2019: 214-218.

[27] NGUYEN T N T, NGUYEN N K, PHAN H, et al. A general network architecture for sound event localization and detection using transfer learning and recurrent neural network[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 935-939.

[28] ZHANG Y, WANG S, LI Z, et al. Data augmentation and class-based ensembled CNN-Conformer networks for sound event localization and detection[R]. [S.l.]: DCASE, 2021.

[29] LEE S H, HWANG J W, SEO S B, et al. Sound event localization and detection using cross-modal attention and parameter sharing for DCASE2021 challenge[R]. [S.l.]: DCASE, 2021.

[30] WU Y, HE K. Group normalization[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018.

[31] RAMACHANDRAN P, ZOPH B, LE Q V. Swish: a self-gated active function[EB/OL]. (2017-10-27)[2024-01-01] https://arxiv.org/abs/1710.05941.

[32] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]. Berlin: Springer, 2018: 3-19.

[33] LIU Y, HOU M, LI A, et al. Automatic detection of timber-cracks in wooden architectural heritage using YOLOv3 algorithm[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020, XLIII-B2-2020: 1471-1476.

[34] POLITIS A, ADAVANNE S, KRAUSE D, et al. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection[EB/OL]. (2021-07-04)[2024-01-01]. https://arxiv.org/sbs/2106.06999v2.

[35] PARK S, JEONG Y, LEE T. Self-attention mechanism for sound event localization and detection[R]. [S.l.]: DCASE, 2021.

Sound event localization and detection network with enhanced feature expression

ZHANG Dongping[1,*], FU Zhentao[1], WANG Zhutao[1], LIN Lili[2], WEI Ming[3]

(1. College of Information Engineering, China Jiliang University, Hangzhou 310018, China;
2. School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China;
3. Hangzhou Aihua Intelligent Technology Co., Ltd., Hangzhou 311121, China)

Abstract: To address the problem that traditional deep learning models are difficult to capture the long-context feature correlations in input feature maps as well as the key feature information in channel and spatial dimensions, resulting in high error rates and unsatisfactory performance in sound event localization and detection (SELD). Based on the baseline model SELDnet in the acoustic scene classification and sound event detection challenge, this paper proposes a feature enhanced sound event localization and detection network (FE-SELDnet). In order to address the issue of function failure to backpropagate, which leads to neuron death, it suggests using group normalization and the SiLU activation function; introducing the convolutional block attention module (CBAM) to capture significant features in both channel and spatial dimensions of acoustic features, suppressing superfluous features, improving network sensitivity and accuracy to feature information, and improving information flow; introducing the Transformer module to capture longer speech context feature association and combine local features to improve the accuracy and robustness of the model in sound event detection and localization tasks. The proposed FE-SELDnet significantly outperforms the original baseline network, according to experimental results on the TUT Sound Events dataset. The error rate decreased from 0.45 to 0.326, the SED and DOA scores decreased from 0.45 and 0.32 to 0.26 and 0.25, respectively, and the F_1 score increased to 79.4%. The algorithm proposed in this paper has higher superiority.

Keywords: sound event localization and detection; enhanced feature expression; attention mechanism; deep learning; group normalization

Received: 2024-01-11; **Accepted**: 2024-02-29; **Published Online**: 2024-03-15 15:16

URL: link.cnki.net/urlid/11.2625.V.20240314.2102.001

Foundation items: Zhejiang Key R & D Project of China (2023C01034,2023C01030,2023C01032)

* **Corresponding author**. E-mail: 06a0303103@cjlu.edu.cn