

基于多尺度注意力融合的声音事件定位与检测

黄翔¹, 杨吉斌¹, 张雄伟¹, 张强¹, 梅鹏程²

- (1. 陆军工程大学 指挥控制工程学院, 江苏 南京 210007;
2. 中国人民解放军 31693 部队, 黑龙江 哈尔滨 150036)

摘要: 深度学习方法已被证明可以大幅度提升声音事件定位与检测(SELDT)性能,而基于CNN的深度SELDT模型往往通过多次卷积和下采样以提取高级特征表示,降低了短时声音事件信号的时间分辨率,空间方位乃至事件类别的估计都会受到影响。为了解决这些问题,提出一种基于多尺度注意力融合(MsAFM)机制的SELDT模型EINV2-MsAFM。该模型基于EINV2框架,计算不同尺度特征的注意力权重以实现不同尺度特征的加权融合。与其他多尺度特征融合方法相比,多尺度注意力融合能够更准确地关注不同尺度条件下事件的重要特征,表现出最佳的融合效果。实验结果表明,相较于基线方法,EINV2-MsAFM模型在TAU 2020数据集上的检测和定位性能表现更优,尤其是在多声源场景下,其检测错误率下降了0.02,定位误差下降了1.1°。

关键词: 深度学习; 声音事件定位与检测; 多尺度注意力融合; 特征融合; 多声源场景

DOI: 10.11907/jjdk.241877

扫描二维码阅读全文:



中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(2026)002-0028-08

Sound Event Localization and Detection Based on Multi-scale Attention Fusion

HUANG Xiang¹, YANG Jibin¹, ZHANG Xiongwei¹, ZHANG Qiang¹, MEI Pengcheng²

- (1. Command and Control Engineering School, Army Engineering University, Nanjing 210007, China;
2. Unit 31693, People's Liberation Army (PLA), Harbin 150036, China)

Abstract: Deep learning methods have been proved to significantly improve the performance of Sound Event Location and Detection (SELDT). However, deep SELDT models based on Convolutional Neural Networks (CNNs) often rely on multiple convolutions and down-sampling to extract high-level feature representation, which blurs the fine information hiding in the signals, especially for those with the short duration, and results in the degraded estimation of location and categories. To solve the problem, this paper proposes a SELDT model based on multi-scale fusion of attention mechanisms (EINV2-MsAFM), which follows the state-of-the-art framework EINV2 and calculates the different scale attention weights to implement the weighted fusion of different scale features. Experimental comparison was conducted on different multiple-scale feature fusion strategies, where the attention-based fusion method (AFM) shows the better fused performance. Thorough simulation experiments results show that the proposed EINV2-MsAFM model performs better in detection and localization on the TAU 2020 dataset compared with the baseline method. In the multi-source scenario, the detection error rate is reduced by 0.02, the localization error is reduced by 1.1°.

Key Words: deep learning; sound event localization and detection; multi-scale attention fusion; feature fusion; multi-source scenario

0 引言

声音事件定位与检测(Sound Event Localization and

Detection, SELDT)通过同时实现声音事件检测(Sound Event Detection, SED)和声源定位(Sound Source Localization, SSL)两个不同属性的任务以形成对场景的完整感知,在环境监测^[1]、语音识别^[2]和智能助听^[3]等方面具有广泛应用。

收稿日期: 2024-12-04

基金项目: 国家自然科学基金项目(62071484); 陆军工程大学基础前沿科技创新项目(KYZYJKQTZQ23001)

作者简介: 黄翔(1999-),女,陆军工程大学指挥控制工程学院硕士研究生,研究方向为人工智能、智能语音处理;杨吉斌(1978-),男,博士,陆军工程大学指挥控制工程学院副教授、硕士生导师,研究方向为模式识别、深度学习、智能语音处理;张雄伟(1965-),男,博士,陆军工程大学指挥控制工程学院教授、博士生导师,研究方向为语音与图像处理、智能语音处理;张强(1991-),男,陆军工程大学指挥控制工程学院博士研究生,研究方向为信号处理、模式识别、深度学习;梅鹏程(1993-),男,硕士,中国人民解放军31693部队助理工程师,研究方向为机器学习、智能语音处理。本文通讯作者: 杨吉斌。

随着机器学习技术的迅速发展,深度学习方法已广泛应用于SELD任务。目前,典型的网络模型有SELDNet^[4]、EINV2^[5-7]等,通过引入多层网络的复杂结构提取信号深层表示特征以提升SELD性能。前者采用三层卷积和GRU单元,后者采用四层卷积和多头自注意力(Multi-Head Self-Attention, MHSA)模块。注意力机制由于实现了对声音信号中关键信息的提取,可以有效提高声音事件定位与检测性能^[8-9]。然而,在特征提取过程中,这些网络使用了多次卷积操作和下采样处理,原始数据的时域分辨率逐渐降低,导致网络在后续处理中难以有效捕获和利用特定时间段内类别、方位中所包含的重要信息。此外,不同类别的声音事件时间持续分布极不均匀,真实场景中存在的一些短时事件在多次卷积和下采样处理后信息丢失严重^[10]。为保证深度特征包含更多的事件细节信息,可以取消部分卷积层的时间维度池化操作,但这往往会带来巨大的计算量,方法适用性不高^[11]。

在典型SELD网络中,由多个卷积模块构成的特征提取网络可以得到不同尺度的特征,这些特征捕获了不同通道、时间和频率尺度上的声音事件信息。对这些多尺度特征进行融合可以改善深度嵌入表示的时间分辨率,以使SELD模型能够更好地理解声音事件的时序和频谱特性,提高对复杂声音环境的感知能力^[12-13]。目前,很多领域都引入了多尺度特征融合方法。Lin等^[14]针对目标检测任务提出特征金字塔网络(Feature Pyramid Networks, FPN),通过建立跨尺度特征金字塔将特征进行相加处理,提高了检测精度和速度。在SED任务上,Wang等^[15]提出一种多尺度全卷积网络,其中引入了特征融合模块,该模块将不同尺度的特征进行了简单拼接处理,但不同尺度的特征可能包含部分重复信息,直接拼接容易导致信息冗余,且增加了模型计算复杂性。在SELD任务中,Sun等^[16]提出一种多尺度特征提取器,通过对特征图通道和空间方向的信息进行加权以增强重要通道的特征,但该提取器只是通过拼接实现了不同尺度特征融合处理。Xie等^[17]设计了一个注意力特征融合单元(Attention Feature Fusion, AFF)以捕获来自不同通道尺度的声源相关特征,但该方法实际上是对特征进行拼接再分割的操作,本质上没有做到对不同通道尺度特征的专注。由于SELD所关注的不同声音事件持续时间不同、频带不同,它们的特征对处理任务没有同等重要性,某些尺度的特征可能对任务的解决更为关键,而其他尺度的特征则可能仅仅提供了一些附加的细节信息。因此,在特征融合时对不同尺度特征的重要性(权重)进行自适应调整,更多地关注相关和有价值的特征,有望提高模型对声音事件的定位和检测能力。基于上述动机,本文提出了一种基于多尺度注意力融合(Multi-scale Attentional Fusion Mechanism, MsAFM)的SELD模型EINV2-MsAFM,通过在模型中引入多尺度特征注意力融合模块,对不同尺度的特征进行加权融合,从而使得模型更多地关

注对SELD任务贡献较大的特征,提高模型性能。

1 多尺度特征融合

在SELD中,对声音信号提取的特征包含时间、频率、空间等不同信息。多尺度特征融合是指对不同尺度的特征进行加权组合,以得到能包含不同时间分辨率、频率分辨率等信息的特征。假设特征张量 X_1 和 X_2 的尺度不同,融合后的结果 y 可以表示为如式(1)所示。

$$y = F(X_1, X_2) \quad (1)$$

其中, F 代表融合方法。根据实现机理不同, F 可以采用传统多尺度特征融合方法和基于注意力的多尺度特征融合方法等。

1.1 传统特征融合方法

传统多尺度特征融合方法一般采用特征相加或拼接的方式实现^[18]。

1.1.1 特征相加方法

特征相加方法使用含有不同时间窗口和频率窗口的卷积神经网络以提取不同时间尺度和频率尺度的特征,如图1所示。此时,利用卷积模块处理不同尺度特征 X_1 和 X_2 ,将它们调整到相同的维度上逐元素相加或相减,得到融合后的特征表示 y ,如式(2)所示。

$$y = C(X_1) + C(X_2) \quad (2)$$

其中, $C(X_1)$ 和 $C(X_2)$ 具有相同维度。

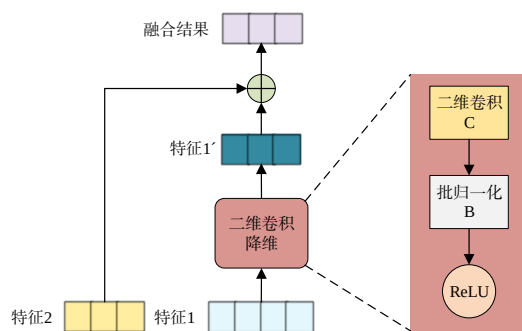


Fig. 1 Multi-scale addition fusion mode structure

图1 多尺度相加融合方式结构

1.1.2 特征拼接方法

拼接方法直接对不同尺度特征进行操作,根据拼接操作的维度不同可将其分为在时间维度上连接和通道维度上连接两种方式。若 X_1 的大小为 $b \times c \times t_1$, X_2 的大小为 $b \times c \times t_2$,则在时间维度上拼接后,特征 y 的大小为 $b \times c \times (t_1 + t_2)$ 。而若 X_1 的大小为 $b \times c_1 \times t$, X_2 的大小为 $b \times c_2 \times t$,则在通道维度上拼接后,特征 y 的大小为 $b \times (c_1 + c_2) \times t$ 。

图2给出了在两个尺度上特征拼接的结构示意图,融合过程可用式(3)表示。

$$S = \text{ReLU}(B(W_c * [X_1, X_2])) \quad (3)$$

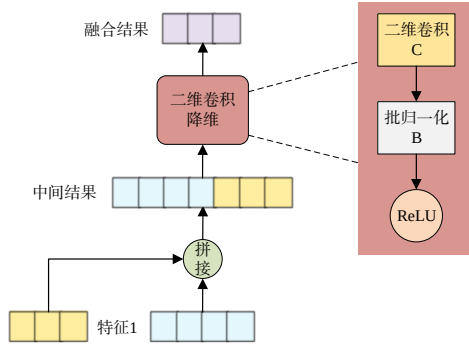


Fig. 2 Multi-scale splicing fusion mode structure

图2 多尺度拼接融合方式结构

其中, W_c 为核大小为 $C \times 1 \times 1$ 卷积滤波器, B 表示批量归一化, S 为最终融合结果。

通过特征拼接,不同尺度的特征可以进行拼接或堆叠,形成一个更复杂、更全面的特征表示,有助于捕捉不同特征之间的关联性和差异性。

相加和拼接方法实现简单,但它们都没有充分考虑不同尺度特征的重要性差异,对于具有不同重要性的特征而言,融合结果并不一定最优。

1.2 基于注意力的多尺度特征融合结构

注意力融合模块 (Attentional Fusion Modules, AFM) 与传统融合方法相比,以两个特征的拼接作为输入,使用两个卷积层计算时频注意力图,最后计算出不同特征的权重,并进行加权^[19]。多尺度注意力融合方式结构如图3所示。

首先将输入的两个特征 X_1 、 X_2 通过降维处理调整到相同大小,注意力图 S 的计算如式(4)所示。

$$S = \tanh(B(W_2 * \text{ReLU}(B(W_1 * [X_1, X_2]))))) \quad (4)$$

其中, W_1 和 W_2 是核大小为 1×1 卷积滤波器, S 具有与 X_1 和 X_2 相同的形状。然后,使用注意力图以融合特征图 X_1 和 X_2 ,如式(5)所示。

$$F = \text{AFM}(X_1, X_2) = (1 + S) \times X_1 + (1 - S) \times X_2 \quad (5)$$

式(5)采用逐元素运算。在 AFM 中, S 表示输入 X_1 和 X_2 之间的相关性。对 S 进行元素加法和减法后, X_1 和 X_2 在融合时权重互补,这使得 AFM 模块能够通过增强一个特征的某个部分,抑制另一个特征的同一部分,即选择每个特

征重要的部分进行融合,从而最大化得到最有用的融合特征。

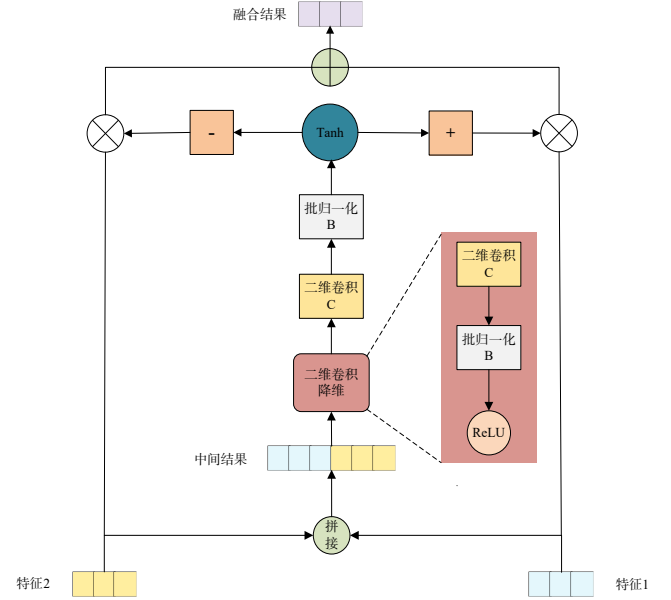


Fig. 3 Multi-scale attention fusion mode structure

图3 多尺度注意力融合方式结构

2 基于多尺度注意力融合的 SELD 模型

2.1 基本模型

本文多尺度注意力融合 SELD 模型采用 EINV2 作为基础框架,称为 EINV2-MsAFM,其模型结构如图4所示。该模型利用 SED 和 DOA 两个分支分别提取高层次特征,两个分支采用多尺度注意力进行不同尺度特征的加权。多尺度注意力特征融合模块采用如图3所示的结构图,可根据不同卷积层之间的组合方式和方向形成不同的组合模式。

在 EINV2-MsAFM 模型中,SED 和 DOA 两个分支上分别堆叠了4次二维卷积模块,在不同时间和频率分辨率下提取音频信号的高层次特征。卷积模块采用批归一化和 ReLU 激活函数以提取声音信号的局部特征信息。在各分支中,使用多级 AFM 模块对不同尺度特征进行融合,以保

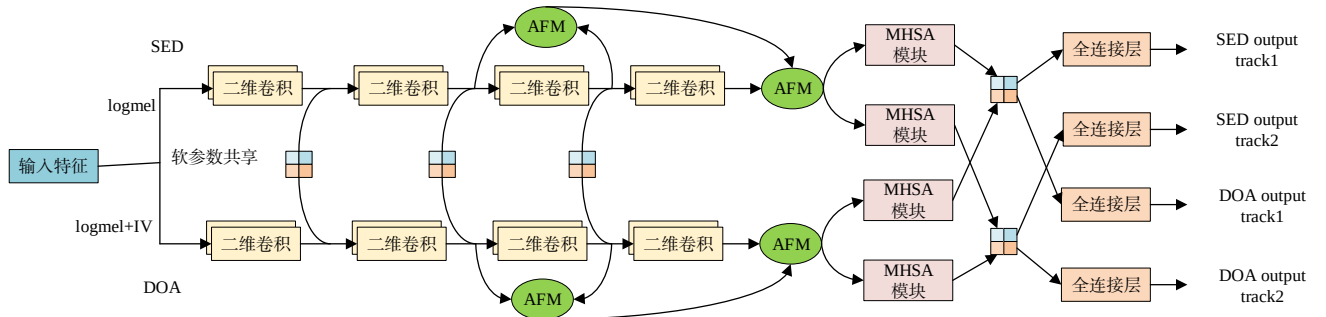


Fig. 4 EINV2-MsAFM model structure

图4 EINV2-MsAFM模型结构

留特征图的关键信息。在高层次特征空间中, 利用 MHSA 模块增强特征捕获细粒度距离相关信息的能力, 有效捕获特征的上下文关联, 最后通过全连接层将深层特征信息进行分类, 并按照 track-wise 的方式输出声学事件的检测和定位结果。

2.2 模型参数设置

SELD 输入是一段多通道音频, 音频通过 STFT 特征提取得到一个四维张量 $[B, C, T, F]$, 其中 B 是批次, C 是通

道数, T 为数据的时间序列长度, F 表示特征频率维度。经过多次二维卷积和多尺度融合后, 特征大小变为 $[B, 512, T/4, F/16]$, 将特征降维到 $[B, 512, T/4]$ 后输入到 MHSA 模块, 输出维度保持不变, 最后将特征大小调整到 $[B, T/4, 512]$ 输入到全连接模块。SED 分支的输出维度为 $[B, 2, N]$, DOA 分支的输出维度为 $[B, 2, 3]$, 其中 N 为声音事件类的总数。EINV2-MsAFM 模型总体结构参数如表 1 所示。

Table 1 EINV2-MsAFM model structural parameters

表 1 EINV2-MsAFM 模型结构参数

网络层	SED 分支		DOA 分支		多尺度融合模块
CONV1	2×(64 个 3×3 的卷积核), 池化 2×2		2×(64 个 3×3 的卷积核), 池化 2×2		
	软参数共享 (64, 2, 2)				
CONV2	2×(128 个 3×3 的卷积核), 池化 2×2		2×(128 个 3×3 的卷积核), 池化 2×2		
	软参数共享 (128, 2, 2)				
CONV3	2×(256 个 3×3 的卷积核), 池化 1×2		2×(256 个 3×3 的卷积核), 池化 1×2		跨 conv2 和 conv3 层融合 256 个 1x1 的卷积核
	软参数共享 (256, 2, 2)				
CONV4	2×(512 个 3×3 的卷积核), 池化 1×2		2×(512 个 3×3 的卷积核), 池化 1×2		跨 conv3 和 conv4 层融合 512 个 1x1 的卷积核
	在频率维度上作平均				
MHSA	2x(d_model=512, h=8)	2x(d_model=512, h=8)	2x(d_model=512, h=8)	2x(d_model=512, h=8)	
	软参数共享 (512, 2, 2)				
FC	512×14, Sigmoid	512×3, Tanh	512×14, Sigmoid	512×3, Tanh	
OUTPUT	SED 轨迹 1	DOA 轨迹 1	SED 轨迹 2	DOA 轨迹 2	

2.3 损失函数

本文使用联合损失函数, SED 分类损失和 DOA 回归损失分别进行加权后进行联合优化, 如式 (6) 所示。

$$loss_{SELD} = \omega loss_{SED} + (1 - \omega) loss_{DOA} \quad (6)$$

其中, ω 是可调整参数, 在实验中对 ω 从 0.3~0.8 进行了优选实验, 最终确定为 0.5。

3 实验与结果分析

3.1 实验设置

3.1.1 数据集

本文在 TAU 2020 数据集上开展验证实验。TAU 2020 数据集包含 14 类声音事件, 共包括 600 条时间长度为 1 min 的录音^[20]。所有音频文件由 15 个室内环境的脉冲响应与不同声音类样本卷积而成, 脉冲响应方向和距离因空间而异。此外, TAU 2020 数据集包含了 30 min 的室外环境噪声记录。数据集采样率为 24 kHz, 分别保存为 FOA 和 MIC 两种 4 通道音频格式。一共分为 6 个交叉验证部分的文件夹, 每个文件夹包含 100 条录音。训练时, 采用前 400 个音频文件进行训练分组, 100 个音频文件进行验证, 最后 100 个音频文件进行测试。这些录音在时空上最多包含两个重叠的声音事件, 这些重叠声音事件可以来自多个方向和距离。

3.1.2 评价指标

本文使用联合性能评估指标, SED 任务采用与位置相关的 F1 得分和错误率 (Error Rate, ER), 即仅考虑在 DOA 距离差小于特定应用阈值时预测的真阳性情况, 阈值一般小于等于 20°。因此, 这两个评价指标被表示为 $ER_{\leq 20^\circ}$ 和 $F1_{\leq 20^\circ}$ 。DOA 任务采用定位误差 (Localization Error, LE_{CD}) 和其相应的定位召回率 (Localization Recall, LR_{CD}) 作为评价指标。只有在类预测结果与给出的评估结果一致时, 才会计算相应的 LE_{CD} 误差和 LR_{CD} 。

$F1_{\leq 20^\circ}$ 指标的计算公式为:

$$F1_{\leq 20^\circ} = \frac{2 \sum_{k=1}^k TP(k)}{2 \sum_{k=1}^k TP(k) + \sum_{k=1}^k FP(k) + \sum_{k=1}^k FN(k)} \quad (7)$$

$ER_{\leq 20^\circ}$ 指标的计算公式为:

$$ER_{\leq 20^\circ} = \frac{\sum_{k=1}^k S(k) + \sum_{k=1}^k D(k) + \sum_{k=1}^k I(k)}{\sum_{k=1}^k N(k)} \quad (8)$$

其中, $N(k)$ 表示声音事件真正活动的事件数, $S(k)$ 表示检测到的声音事件类中预测错误的事件数量, $I(k)$ 和 $D(k)$ 分别表示插入和删除的种类数量。它们可分别通过式 (9)~式 (11) 加以表示。

$$S(k) = \min(FN(k), FP(k)) \quad (9)$$

$$I(k) = \max(0, FP(k) - FN(k)) \quad (10)$$

$$D(k) = \max(0, FN(k) - FP(k)) \quad (11)$$

LR_{CD} 指标和 LE_{CD} 的计算公式如式 (12)、式 (13) 所示。

$$LR_{CD} = \frac{TP}{TP + FN} \quad (12)$$

$$LE_{CD} = \arccos(u_{ref} \cdot u_{pre}) \quad (13)$$

其中, u_{ref} 表示真实标签对应的笛卡尔向量值, u_{pre} 代表预测出来的笛卡尔向量值。

SELD 分数是 SELD 任务的整体性能指标, 在模型评估时, SELD 分数越低, 代表模型性能越好, 计算公式如式 (14) 所示。

$$SELD \text{ 分数} = \frac{ER_{\leq 20^\circ} + (1 - F_{\leq 20^\circ}) + \frac{LE_{CD}}{180^\circ} + (1 - LR_{CD})}{4} \times 100\% \quad (14)$$

3.1.3 模型实现

在验证实验中, 网络模型基于 PyTorch 框架实现, 采用 EINV2 作为基线方法。为了确保比较的公平性, EINV2-MsAFM 模型的训练参数保持与基线方法相同。在实验前 120 轮, 学习率设置为 1×10^{-3} , 在最后 30 轮, 学习率设置为 1×10^{-4} 。模型使用 Adam 优化器默认参数进行损失优化。验证实验在配备 NVIDIA 2080Ti GPU 的工作站上开展。

3.2 不同多尺度融合方案性能比较

为探索多尺度注意力融合结构在 SELD 任务中所发挥的作用, 分别对不同多尺度融合方案进行了对比实验, 并对 AFF 融合方法^[17]进行对比实验, 实验结果如表 2 所示。多尺度相加模型记为 EINV2-相加, 多尺度拼接模型记为 EINV2-拼接, 与 EINV2-MsAFM 一样, EINV2-相加和 EINV2-拼接使用同样的卷积模块进行融合。最优指标得分用黑体表示。

Table 2 Performance comparison of different fusion schemes in

EINV2 model

表 2 不同融合方案在 EINV2 上的性能比较

模型	$ER_{\leq 20^\circ} \downarrow$	$F1_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow / ^\circ$	$LR_{CD} \uparrow$	SELD 分数 $\downarrow$
EINV2	0.34	0.75	9.61	0.80	0.213
EINV2-MsAFM	0.32	0.76	8.90	0.80	0.205
EINV2-相加	0.33	0.75	9.12	0.79	0.211
EINV2-拼接	0.43	0.68	10.66	0.74	0.268
EINV2-MFF	0.66	0.57	10.78	0.64	0.377

从表 2 结果可以看出, 在不同多尺度特征融合方案中, EINV2-MsAFM 模型和 EINV2-相加模型相较于原模型在绝大部分指标上都获得了更好表现, LE_{CD} 指标和 SELD 分数表现更佳, LR_{CD} 指标持平。其中, EINV2-MsAFM 模型效果更好, 这是因为融合不同尺度的特征信息可以丰富全局特征和局部特征之间的关系, 获得更全面的特征信息。不同尺度的特征信息对于任务的重要性不同, 简单的相加操作可能导致不同尺度特征之间的平衡问题。而采用多尺度拼接方法时, 效果比原模型各指标都要差, 这是因为多尺度拼接是将来自不同尺度的特征在通道维度上直接拼接在一起, 形成一个通道维度更宽的特征向量, 再通过反卷积将通道维度进行降维。这种方式实际上是通过卷

积核给每个通道赋予权重, 虽然可以保证多尺度信息维度不变, 但由于反卷积压缩了特征表示的维度, 导致部分重要信息丢失。

同时, 从表 2 结果可以看出, 本文模型在各指标上都优于 EINV2-MFF 模型。EINV2-MFF 模型的检测错误率明显高出很多, 这是因为在 MFF 方案中仅仅将两个不同尺度的特征进行几次卷积处理和拼接处理, 并没有对两个特征图的重要性进行权衡, 因而融合的特征结果并不利于模型最终性能提升。

3.3 EINV2-MsAFM 模型不同实现结构比较

EINV2 模型包含了 4 层卷积层, 为了研究网络不同阶段(层)的特征对 EINV2-MsAFM 模型的影响, 分别对不同卷积层的特征进行了多尺度特征聚合, 结果如表 3 所示。其中, EINV2-MsAFM_LXXX 代表采用相应卷积层特征进行多尺度特征融合的方案。

Table 3 Performance comparison of different convolution layers fused in EINV2-MsAFM model

表 3 不同卷积层融合在 EINV2-MsAFM 模型上的性能比较

模型	$ER_{\leq 20^\circ} \downarrow$	$F1_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow / ^\circ$	$LR_{CD} \uparrow$	SELD 分数 $\downarrow$
EINV2	0.34	0.75	9.61	0.80	0.213
EINV2-MsAFM_L1234	0.35	0.73	9.34	0.78	0.233
EINV2-MsAFM_L234	0.32	0.76	8.90	0.80	0.205
EINV2-MsAFM_L34	0.34	0.75	8.90	0.79	0.214

从表 3 结果可以看出, 采用全部 4 层特征进行融合的林 INV2-MsAFM_L1234 模型在检测错误率和 F1 分数方面略低于 EINV2 模型, 但在定位误差方面有所改善。采用第二、三、四层特征进行融合的林 INV2-MsAFM_L234 模型和仅采用第三、四层特征融合的林 INV2-MsAFM_L34 模型在定位误差 LR_{CD} 和 SELD 分数方面表现较好, EINV2-MsAFM_L234 模型具有更低的定位误差和 SELD 分数。这是由于不同层级的特征在表达能力、重要性和互补性方面存在差异。在较低层级的网络层中, 由于特征映射较浅, 特征在保留原始信息的同时可能包含一些冗余信息或者噪声, 这种冗余或噪声会对特征融合产生负面影响; 而高层的卷积特征具有更丰富的语义特征, 能够更好地捕捉声源的位置信息。因此, 选择第二、三、四层进行特征进行融合更能精确地推断声源类别和位置, 从而提高模型性能。

同时, 多尺度特征的融合方式通常包括自底而上和自顶而下两种方式, 两种结构如图 5 所示。自底而上的聚合方式有助于保留一些原始特征信息, 防止信息丢失; 自顶而下的聚合方式有助于抑制底层中的噪声信息。为确定多尺度特征融合的最佳方式, 对自底而上 EINV2-MsAFM_u 和自顶而下 EINV2-MsAFM_d 这两种聚合方式和双边聚合方案 EINV2-MsAFM_b 的性能进行了比较分析。3 个方案都使用 2 个 AFM 模块, 各指标结果如表 4 所示。

从表 4 结果可以看出, 相较于 EINV2, EINV2-

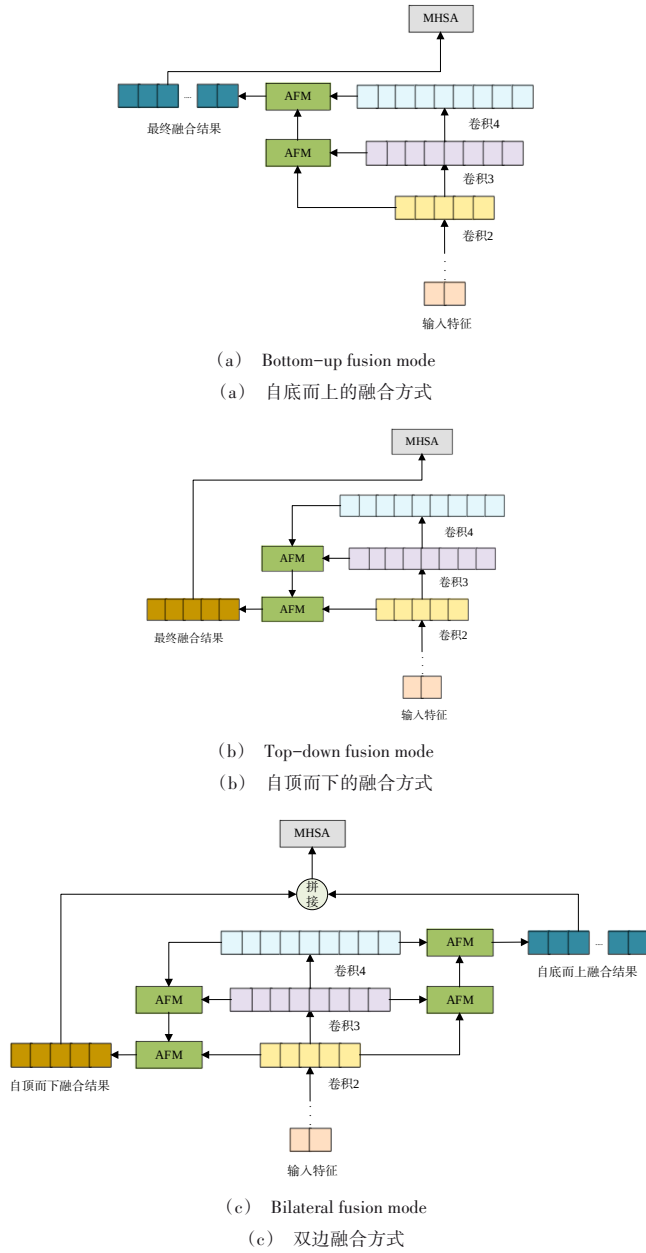


Fig. 5 Diagram of different fusion modes of multi-scale features

图 5 多尺度特征不同融合方式示意图

Table 4 Performance comparison of top-down and bottom-up fusion modes in EINV2-MsAFM model

表 4 自顶而下和自底而上聚合方式在 EINV2-MsAFM 模型上的性能比较

模型	$ER_{\leq 20^\circ} \downarrow$	$F1_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow / ^\circ$	$LR_{CD} \uparrow$	SELD 分数 $\downarrow$
EINV2	0.34	0.75	9.61	0.80	0.213
EINV2-MsAFM_u	0.32	0.76	8.90	0.80	0.205
EINV2-MsAFM_d	0.42	0.68	10.71	0.74	0.267
EINV2-MsAFM_b	0.73	0.37	16.83	0.46	0.498

M_sAFM_u 模型在各指标上都表现很好, 这表明自底而上的聚合方式可以通过将底层特征所包含的细节和局部信息与高层特征进行融合, 使得模型能够更好地捕捉到声音的空间位置和细节特征。而 EINV2-M_sAFM_d 模型在各项指标上表现较差, 可以推断, 这是在抑制底层噪声的同时

对底层特征中的细节信息产生了抑制。当高层特征与底层特征进行融合时, 高层特征可能会抑制底层特征中的细节信息, 导致模型性能不如预期。EINV2-M_sAFM_b 模型在这组实验中表现最差, 这是由于融合过程中同时保留了有用信息和噪声信息。

3.4 不同模型在 TAU 2020 数据集上的性能分析

为了验证本文 EINV2-M_sAFM 模型的有效性, 将其与 CECANet^[8]、MVANet^[9] 和 EINV2-CMA^[21] 等先进的 SELD 模型进行了对比实验。不同 SELD 模型在 TAU 2020 数据集上的实验结果如表 5 所示。

Table 5 Performance comparison of different models on TAU2020

dataset

表 5 不同模型在 TAU2020 数据集上的性能比较

模型	$ER_{\leq 20^\circ} \downarrow$	$F1_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow / ^\circ$	$LR_{CD} \uparrow$	SELD 分数 $\downarrow$
EINV2-MsAFM	0.32	0.76	8.90	0.80	0.205
CECANet	0.33	0.76	9.32	0.80	0.206
MVANet	0.32	0.76	8.74	0.80	0.203
EINV2-CMA	0.36	0.71	12.2	0.86	0.231

从表 5 结果可以看出, EINV2-M_sAFM 模型相较于 CECANet 和 EINV2-CMA 模型, $ER_{\leq 20^\circ}$ 指标和 LE_{CD} 指标明显下降, 获得了良好效果。与 MVANet 模型相比, 其 LE_{CD} 指标和 SELD 分数高出 0.16 和 0.002, 表现略差, 其他指标持平。实验表明, EINV2-M_sAFM 模型对 SELD 的定位和检测性能改善明显, 效果较好。

3.5 不同场景数据的性能分析

为了测试模型在不同场景数据集上的效果, 分别在包含重叠源的 TAU 2020 多混响数据集和包含干扰事件的 TAU 2021 数据集^[22] 上进行了实验, 并利用单声源、多声源测试数据对 EINV2 模型和 EINV2-M_sAFM 进行了测试, 实验结果如表 6 所示。

从表 6 结果可以看出, EINV2-M_sAFM 模型在不同场景数据集上, 无论是单声源数据集还是多声源数据集上的性能都明显优于 EINV2 模型。尤其是在多声源场景下, EINV2-M_sAFM 模型的 $ER_{\leq 20^\circ}$ 指标在 TAU 2020 和 TAU 2021 数据集上较 EINV2 模型下降了 0.02、0.07, LE_{CD} 指标下降了 1.1°、1.79°, SELD 分数则下降了 0.013 和 0.03。这说明采用多尺度注意力融合模块, 能够有效提高重叠源的定位和检测性能。

3.6 定位与检测结果分析

为了更好地展示模型加入多尺度融合注意力方案所带来的效果, 对测试结果进行可视化分析。图 6、图 7 是 TAU 2020 数据集中 mix017 和 mix012 两个音频文件在 EINV2 模型和 EINV2-M_sAFM 模型上的可视化结果比较。mix017 是一段单声源的音频, mix012 是包含多事件重叠源的音频。不同声事件用不同颜色表示, 如绿色表示钢琴声, 红色表示婴儿哭泣声, 粉色表示男性演讲声, 蓝色表示烧火声, 紫色表示警报声等。

Table 6 Performance comparison of EINV2 and EINV2-MsAFM models on different scenario datasets

表 6 EINV2 和 EINV2-MsAFM 模型在不同场景数据集上的性能比较

模型	数据集	声源数量	$ER_{\leq 20^\circ} \downarrow$	$F1_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow / ^\circ$	$LR_{CD} \uparrow$	SELD 分数 $\downarrow$
EINV2	TAU 2020	单声源	0.27	0.80	6.57	0.79	0.169
		多声源	0.45	0.65	11.10	0.72	0.286
EINV2	TAU 2021	单声源	0.36	0.72	9.10	0.71	0.245
		多声源	0.55	0.58	14.92	0.64	0.353
EINV2-MsAFM	TAU 2020	单声源	0.25	0.80	6.42	0.80	0.162
		多声源	0.43	0.67	10.00	0.73	0.273
EINV2-MsAFM	TAU 2021	单声源	0.33	0.72	8.93	0.73	0.232
		多声源	0.48	0.60	13.13	0.66	0.323

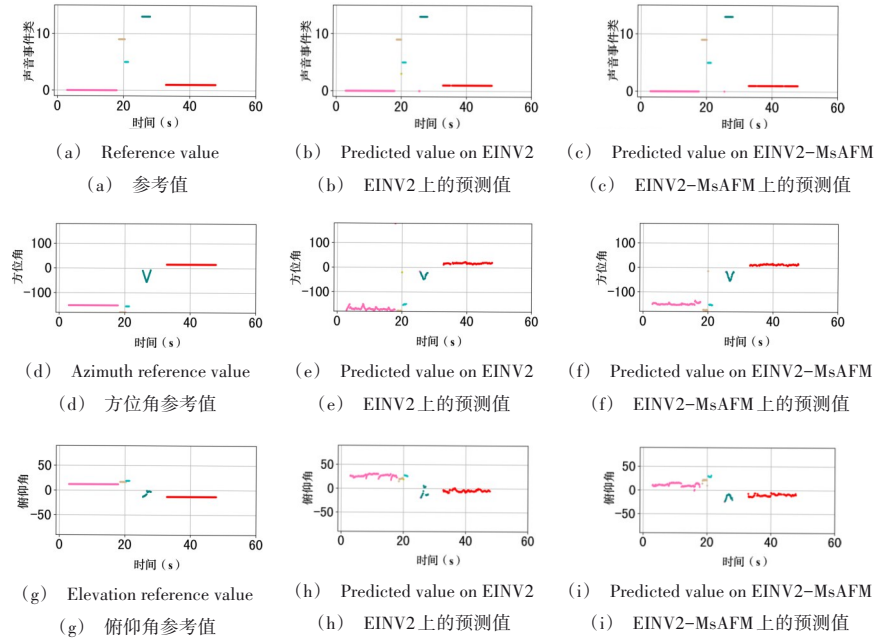


Fig. 6 Comparison of EINV2 and EINV2-MsAFM's reference and predicted values of detection, azimuth and elevation on mix017 data

图 6 EINV2 与 EINV2-MsAFM 在 mix017 数据上的检测、方位角和俯仰角参考与预测值比较

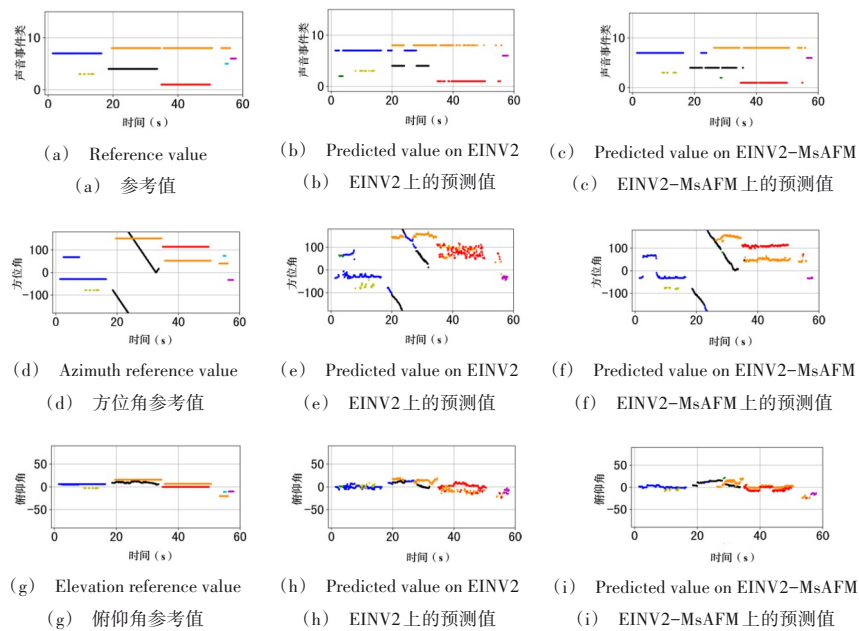


Fig. 7 Comparison of EINV2 and EINV2-MsAFM's reference and predicted values of detection, azimuth and elevation on mix012 data

图 7 EINV2 与 EINV2-MsAFM 在 mix012 数据上的检测、方位角和俯仰角参考与预测值比较

图6给出了在ov1条件下的估计结果,可以看出,EINV2-MsAFM模型相较于EINV2,DOA估计的轨迹更加清晰,结果更加精确,尤其在第0~20 s内更为明显。图7给出了多声源数据估计结果,可以看出,在第35~50 s内,EINV2-MsAFM模型对警报声和婴儿哭泣声的类别判定、位置估计更为准确。

4 结语

本文提出了基于多尺度注意力融合的SELD模型EINV2-MsAFM,通过引入基于注意力的特征融合方法,将不同尺度的特征按照其重要性进行加权融合,解决了已有模型存在的多尺度特征利用不足等问题。实验结果表明,该模型在检测错误率和定位误差上性能均得到了改善,并在单声源和多声源场景下都取得了良好效果。尤其是在多声源场景下,EINV2-MsAFM模型在TAU 2020数据集上的表现更优,其检测错误率下降了0.02,定位误差下降了 1.1° 。未来,将进一步研究在不同的场景下,优化多尺度注意力融合机制设计,以进一步提升强噪声和混响条件下敏感事件的检测和定位性能。

参考文献:

- [1] ZHANG L, LIN Z Y, LI F D, et al. Experimental system design for industrial equipment status monitoring based on environmental sound recognition [J]. *Laboratory Research and Exploration*, 2024, 43(8): 47-51.
张雷,林子煜,李飞达,等.基于环境声识别的工业设备状态监测实验系统设计[J]. *实验室研究与探索*, 2024, 43(8): 47-51.
- [2] SMITH J. Hidden emotion detection from speech and text using deep learning approach [J]. *Journal of Deep Learning*, 2024, 10(2): 234-245.
- [3] LI M. Research on intelligent inspection decision of inspection robot in substation based on artificial intelligence [J]. *Electrical Technology and Economy*, 2024(9): 1-4.
李铭.基于人工智能的变电所巡检机器人智能巡检决策研究[J]. *电气技术与经济*, 2024(9): 1-4.
- [4] 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*, 2018, 13(1): 34-48.
- [5] CAO Y, IQBAL T, KONG Q, et al. An improved event-independent network for polyphonic sound event localization and detection [C]//*IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021: 885-889.
- [6] HU J, CAO Y, WU M, et al. A track-wise ensemble event independent network for polyphonic sound event localization and detection [C]//*2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: 9196-9200.
- [7] HU J, CAO Y, WU M, et al. Sound event localization and detection for real spatial sound scenes: event-independent network and data augmentation chains [DB/OL]. <https://arxiv.org/abs/2209.01802>.
- [8] XU C D, LIU H, MIN Y, et al. Sound event localization and detection based on dual attention [J]. *Computer Engineering and Applications*, 2023, 59(19): 99-105.
许春冬,刘昊,闵源,等.基于双重注意力的声音事件定位与检测[J]. *计算机工程与应用*, 2023, 59(19): 99-105.
- [9] YANG J B, HUANG X, ZHANG X W, et al. Sound event localization and detection model based on multi-view attentions [J]. *Signal Processing*, 2024, 40(2): 385-395.
杨吉斌,黄翔,张雄伟,等.采用多视角注意力的声音事件定位与检测[J]. *信号处理*, 2024, 40(2): 385-395.
- [10] CHENG S, DU J, WANG Q, et al. Improving sound event localization and detection with class-dependent sound separation for real-world scenarios [C]//*Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023: 2068-2073.
- [11] WANG Q, WU H, JING Z, et al. A model ensemble approach for sound event localization and detection [C]//*12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021: 1-5.
- [12] HEALY B, MCNAMEE P, AHMADABADI Z N. Feature aggregation in joint sound classification and localization neural networks [J]. *IEEE Access*, 2024, 12: 109157-109170.
- [13] ZHANG S, WANG K, LI J, et al. Resnet-conformer network using multi-scale channel attention for sound event localization and detection in real scenes [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, 32(4): 1251-1264.
- [14] LIN T, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 2117-2125.
- [15] WANG Y, ZHAO G, XIONG K, et al. Multi-scale and single-scale fully convolutional networks for sound event detection [J]. *Neurocomputing*, 2021, 421: 51-65.
- [16] SUN X, ZHU X, HU Y, et al. Sound event localization and detection based on CRNN using adaptive hybrid convolution and multi-scale feature extractor [R]. *DCASE Technical Report*, 2021.
- [17] XIE Y, LIU J, HU Y. A polyphonic SELD network based on attentive feature fusion and multi-stage training strategy [C]//*4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, 2023: 650-654.
- [18] ZHANG M, ZHAO Y, CHEN R, et al. Remote sensing collaborative classification using multimodal adaptive modulation network [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-12.
- [19] QI J, GUO W, GU B. Bidirectional multiscale feature aggregation for speaker verification [C]//*Interspeech*, 2021: 71-75.
- [20] POLITIS A, ADAVANNE S, VIRTANEN T. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection [DB/OL]. <https://arxiv.org/abs/2006.01919>, 2020.
- [21] LEE S H, HWANG J W, SEO S, et al. Sound event localization and detection using cross-model attention and parameter sharing for DCASE2021 Challenge [R]. *Technical Report*, 2021.
- [22] POLITIS A, ADAVANNE S, KRAUSE D A, et al. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection [DB/OL]. <https://arxiv.org/abs/2106.06999>, 2021.

(责任编辑:孙娟)