基于多任务学习与动态卷积的帧级别小样本声音事件检测方法

方昕[1,2]，赵鹏源[3]，张雨涛[3]，严根伟[1]，闫祖龙[1]，邹亮[3]

（1.科大讯飞股份有限公司讯飞研究院，合肥 230088；2.中国科学技术大学信息科学技术学院，合肥230026 ; 3.中国矿业大学信息与控制工程学院，徐州 221116）

摘 要：声音事件检测（Sound Event Detection, SED）作为智能感知与环境交互的关键技术，广泛应用于工业监测、生态保护和安防预警等领域。为应对SED任务中标注样本稀缺、事件时间分辨率不足等问题，本文提出了一种融合多任务学习与动态卷积的帧级别小样本声音事件检测方法。该方法在帧级别检测模型的基础上引入前景/背景声音二分类辅助任务，结合基于状态空间建模的NetMamba编码器与频率动态卷积结构（Frequency Dynamic Convolution, FDC）提升模型的长时序建模能力与频域特征自适应提取能力。同时，设计了一种线性时域扰动的数据增强策略TimeFilterAug，以模拟复杂声学环境下的噪声干扰，增强模型的泛化能力。所提方法在DCASE 2024国际挑战赛中取得56.7%的F1分数，排名第二。实验结果表明，该框架在短时瞬态事件识别及噪声鲁棒性方面具有显著优势，具备良好的工程应用前景。

关键词：多任务学习；帧级别框架；NetMamba编码器；滤波增强；动态卷积

Frame-Level Few-Shot Sound Event Detection Framework Based on Multi-Task Learning and Dynamic Convolution

FANG Xin[1,2], ZHAO Pengyuan[3], ZHANG Yutao[3], YAN Genwei[1], YAN Zulong[1], ZOU Liang[3]

(1.iFlytek Research, iFLYTEK Co., Ltd., Hefei 230088, China; 2.School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China; 3.School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China）

Abstract：Sound Event Detection (SED) plays a critical role in intelligent perception systems, with broad applications in industrial monitoring, ecological surveillance, and public safety. To address the limitations of low-resource SED scenarios, such as insufficient labeled data and limited temporal resolution, this paper proposes a frame-level few-shot sound event detection method based on multi-task learning and dynamic convolution. The proposed method integrates a foreground/background sound classification (FBSC) task alongside the primary SED objective, and combines a NetMamba encoder with a Frequency Dynamic Convolution (FDC) module to enhance long-range temporal modeling and frequency-adaptive feature extraction. Additionally, a linear time-domain perturbation strategy, TimeFilterAug, is designed to simulate noise interference in complex acoustic environments, further improving the model's generalization capability. The proposed method achieved an F1-score of 56.7% in the DCASE 2024 Challenge, ranking second. Experimental results confirm its effectiveness in detecting transient events and

基金项目：国家自然科学基金（编号：62373368，62473360）

maintaining robustness under noisy conditions, demonstrating strong potential for practical deployment.

Key words: Multi-task Learning; Frame-level framework; NetMamba Encoder; TimeFilterAug; Dynamic convolution.

引 言

声音事件检测（Sound Event Detection, SED）旨在从连续音频中识别特定声音事件并标注起止时刻，可从复杂多源环境声中提取关键信息，支撑系统感知环境变化与识别异常状态。随着人工智能尤其是深度学习与智能音频信号处理技术的快速发展，SED已广泛应用于工业设备故障监测[1]、生态环境保护[2]、医疗健康监护[3]和公共安全管理[4]等领域。高效、准确的声音事件检测不仅提升了系统在动态声学环境中的感知精度与响应效率，也为构建智能化、自动化的多模态感知系统提供了重要支撑。

当前主流SED方法多基于卷积神经网络（Convolutional Neural Networks, CNN）与卷积递归结构（Convolutional Recurrent Neural Network, CRNN）[5]，借助深层特征提取与时序建模能力在多源事件识别中表现出优异性能。然而，这类方法普遍依赖大规模、高质量的标注样本，对模型训练提出了较高的数据需求。声音数据的采集与标注过程通常成本高昂且耗时费力[6]，在实际应用中，尤其是工业设备状态监测[7]、稀有物种检测等场景，难以获取充足且多样化的样本，导致模型在小样本条件下的检测性能显著下降。为此，面向小样本的声音事件检测（Few-Shot Sound Event Detection, FSSED）成为近年来SED研究的重要方向，旨在通过极少量支持样本实现新类别声音事件的有效识别[8]。

针对小样本声音事件检测，已有研究主要沿两类技术路径展开。一类是基于模板匹配的方法，通过计算支持样本与查询样本在特征空间的相似性度量完成分类，典型代表如基于归一化特征进行交叉相关性计算的匹配策略[9]。该类方法结构简单，推理高效，但对特征表征质量敏感，在复杂噪声或多事件重叠背景下易出现性能退化。另一类是基于原型网络（Prototypical Networks, ProtoNet）的方法，通过构建类别原型表示并基于最近邻规则完成分类，同时引入对比学习、转导推理等策略以增强模型泛化能力[10][11][12][13]。尽管这些方法在特定场景下展现出良好的小样本适应性，但普遍存在时序粒度粗糙、复杂环境鲁棒性不足的问题，难以满足实际应用需求，具体包括：（1）时序粒度不足。现有方法多采用固定段级（Segment-Level）预测，难以捕捉短时、瞬态声音事件的细粒度动态特征，导致事件边界检测不准确，影响整体检测性能[14]。深层特征建模能力受限。多数方法侧重浅层特征匹配或简单度量学习，忽视了上下文信息建模与长时依赖关系捕捉，导致复杂声学环境下的事件识别能力不足[15]。（2）噪声环境适应性差。在现实应用中，声音信号常伴随远场噪声、环境脉冲干扰及多源事件叠加，现有方法在面对声学场景变化时鲁棒性较差，易出现性能退化[16][17]。（3）为克服上述挑战，提升小样本条件下的帧级别检测精度与复杂场景适应性，本文提出一种融合多任务学习与频率动态卷积建模的帧级别小样本声音事件检测框架，全面提升模型在粒度分辨率、特征表达力与噪声适应性方面的能力，主要贡献如下：

（1） 提出融合前景/背景声音二分类（Foreground-Background Sound Classification, FBSC）辅助任务的帧级别检测框架，有效提升短时瞬态事件的检测精度，并增强复杂噪声环境下的事件判别能力。

（2） 引入基于状态空间建模的NetMamba编码器与设计频率动态卷积（Frequency Dynamic Convolution, FDC）特征提取模块，联合提升模型在长时依赖建模与频率敏感特征捕捉方面的能力，增强复杂声学环境下的特征表达与建模效果。

（3） 提出TimeFilterAug数据增强策略，通过模拟线性时间扰动形式的动态背景噪声变化，显著提升模型在多噪声、多场景环境下的小样本迁移能力与鲁棒性。

本文方法在DCASE 2024国际挑战赛Task5少样本生物声学事件检测任务中进行了系统验证[18]，在官方评估集上取得了56.7%的F1分数，最终排名第二。实验结果表明，该方法在小样本帧级别检测、短时事件识别与复杂环境适应方面具备显著优势，具有良好的实际部署价值与学术研究前景。

1 数据集与预处理

1.1 数据集描述

本研究采用了来自DCASE2024小样本生物声学事件检测任务的数据集[9]，包括开发集 X_{dev} 和评估集 X_{test} 。其中，开发集 X_{dev} 被预先划分为训练集 X_{train} 和验证集 X_{val} ，并确保二者之间无重叠，即 $X_{train} \cap X_{\mathrm{val}} = \varnothing$ ，训练集 X_{train} 由BV(BirdVox)、HT(Hyenas)、JD(Jackdaws)、MT(Meerkats)以及WMW(WildMammalWatch)五个子数据集组成，分别来自不同生物声学数据源，包含174个音频文件，共21小时的音频时长，跨越47个生物声学类。每个音频文件均附带多类标注，标签包括目标事件（POS）、背景声音（NEG）和不确定类（UNK）。由于WMW中存在多处错误标注信息，剔除对应音频文件后，训练集共包含约15小时的音频时长，覆盖40个生物声学类别。验证集 X_{val} 由6个独立子数据集构成，共包含43个音频文件，约50小时的音频时长，跨越9个生物声学类。评估集由66个音频片段组成，覆盖8个不同的数据源子集。这些音频采集自全球多个生态区域，涵盖不同的物种和采集设备，可进一步检验模型在未知场景下的迁移能力与稳健性。上述声音检测数据集的基本信息如表1所示，其中包括：数据集的名称、事件个数、总类别数、总事件个数、总时长、事件平均时长。

表1 实验数据基本信息

Table1 Basic information of experimental datasets

数据集名称		事件个数	类别数	总时长	事件平均时长（秒）
X_{train}	BV	9026	11	10 h	0.15
	HT	611	5	5 h	1.42
	JD	357	1	10min	0.12
	MT	1294	4	1h10min	0.14
	WMW	2644	19	2h40 min	1.54
X_{val}	HB	712	1	2h38 min	11.6
	ME	73	2	20 min	0.19
	PB	292	2	3 h	0.11
	PB24	350	2	2h	0.08
	RD	1372	1	18h	1.52
	PW	705	1	24h	2.21

验证集与评估集中的每个音频文件仅对应一个特定的目标类别，均视为一个独立的小样本检测任务。如图1所示，

基金项目：国家自然科学基金（编号：62373368，62473360）

每个任务的输入音频划分为两个部分：支持集（Support Set）与查询集（Query Set）。支持集由前五个目标事件片段（POS）及其相邻的背景片段（NEG）交替组成，记作 $S=\{N_1,P_1,\cdots,N_5,P_5\}$，其中 P_i 和 N_i 分别代表第 i 个带有逐帧标签的目标事件与背景片段。基于输入音频支持集中的POS与NEG 帧级别标签片段，通过帧级别小样本声音事件检测系统实现对查询集中连续帧的逐帧分类，从而精准定位后续查询集中内的目标事件。

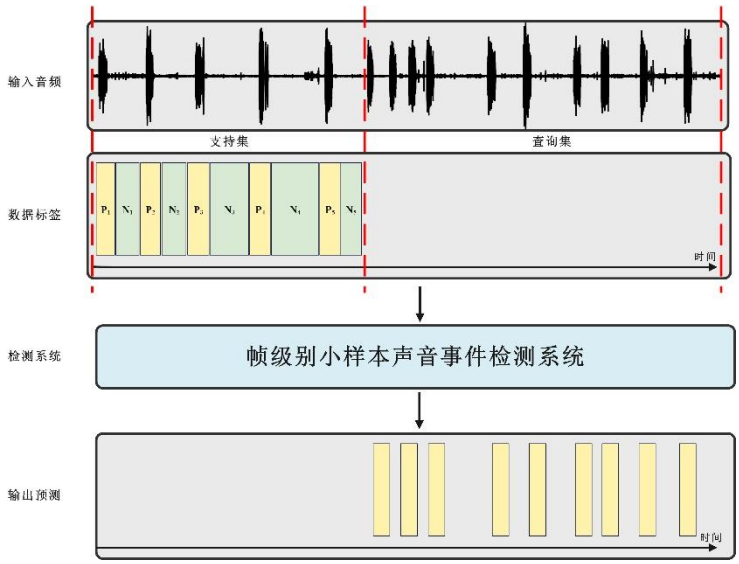


图1 声音事件检测数据结构可视化

Fig.1 Visualization of Data Structure for Sound Event Detection

1.2 帧级别预处理策略

在数据预处理阶段，为缓解非同源录制条件下采样率差异对频谱分辨率的影响，本文统计了训练集中音频采样率的中位数，并统一将所有音频重采样至22050 Hz。重采样后，对音频信号进行短时分帧与加窗处理，以获得近似平稳的帧级别表示。随后，对音频帧序列应用长度为5秒、帧移为1秒（对应4秒重叠）的滑动窗口，对应帧级别参数为431帧长和86帧窗移动，以确保目标事件在时间维度上的连续性与覆盖度。在此基础上，进一步提取每个窗口对应的对数梅尔频谱图（Log-Mel Spectrogram）[19]，以获取稳定且具有区分性的时频域声学特征。

在标签对齐过程中，根据采样率与帧长信息将事件标注时间戳转换为帧级别标签，未对应任何事件的帧标记为“0”类（背景类），以充分利用背景音中的上下文信息，增强模型对非事件帧的判别能力。此外，为增强训练样本的多样性，在基类数据训练阶段，本文采用Kaldi工具对音频进行语速变换（0.9、1.0、1.1倍速率），以扩充数据规模提升模型鲁棒性。

2 多任务训练框架

小样本声音事件检测在存在稀疏标注、复杂声学干扰和短时瞬态事件等现实挑战，常使传统帧级别检测方法在特征表征、时序建模和环境适应性方面表现受限[16]，模型常出现漏检或边界定位不准，显著降低预测精度。为了解决这一问题，本文提出一种多任务训练框架，集成频率自适应建模、前/背景辅助任

务与高效时序建模机制，增强模型在复杂噪声环境下的声音事件检测能力，具体如图2所示。

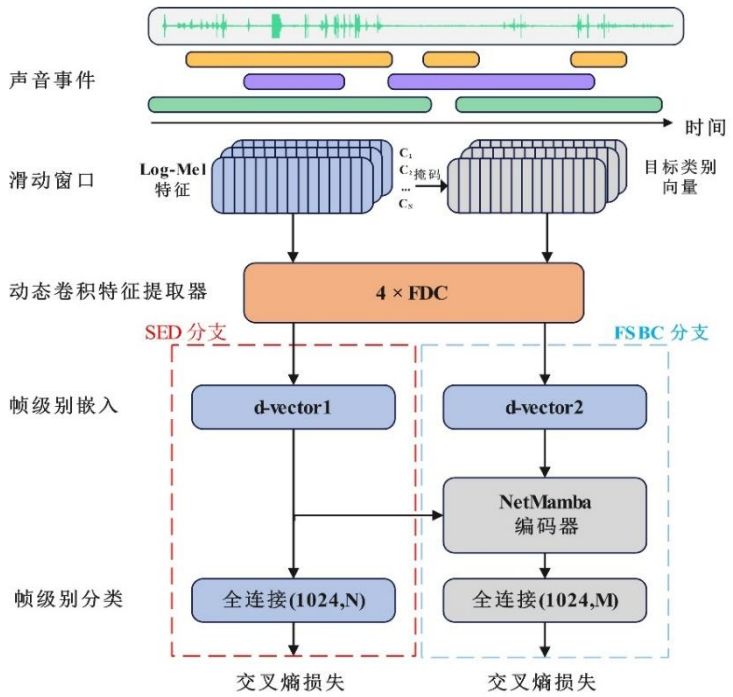


图2　多任务训练框架

Fig.2　Multitask training framework

首先采用帧级别预处理方式对输入音频进行音频特征提取与帧级别样本窗的切分，得到预处理完成的Log-Mel特征。在SED分支中，生成的Log-Mel特征经由FDC模块进行处理，对每个窗口提取高维音频特征表示，并将其编码为1024维度帧级别嵌入向量d-vector1，通过全连接层解码进行逐帧分类，以实现对40类声音事件的帧级别定位预测。在FBSC分支中，采用迭代式的前/后景动态划分机制，生成与Log-Mel特征对齐的类别掩码，引导模型对各类别施加均衡注意力，缓解声音事件类别不均引发的注意力偏置。具体地，对于每个滑动窗口，先统计其中所有出现的事件类别，并在每轮训练中轮换选定其中一个作为前景类，其余视为背景类。随后，根据切分窗中每一帧的真实标签，保留属于目标前景类的帧级别特征，其余帧置零，从而构建目标类别特征掩码。将包含目标事件帧的特征子集经FDC编码器生成 1024 维帧级别向量d-vector2，并与SED分支输出的d-vector1 在通道维度拼接后输入 NetMamba 编码器进行时序建模，最终通过全连接层映射为帧级别前景/背景概率。前/后景动态划分机制可表示为：

$$x_i = \begin{cases} x_i & if\ y_i = c_t \\ 0 & else \end{cases} \tag{1}$$

其中 x_i 表示LogMel特征的一个帧， y_i 是其对应的标签，是选定的目标类别。

在训练阶段，多任务检测框架同时计算两个子任务的损失，损失函数在训练过程中联合优化模型参数，从而提升模型在复杂环境中对长短时事件的联合建模能力和检测的鲁棒性，两个分支中均应用了交叉熵（Cross entropy， CE）损失，分别记作 l_1 和 l_2 其表达式如下

基金项目：国家自然科学基金（编号：62373368，62473360）

$$l_1 = -\frac{1}{L}\sum_{i=1}^{l}\sum_{j=1}^{c}((Y_i == j) \odot M_i)\log(f_{\phi 1}(X_i)) \tag{2}$$

$$l_2 = -\frac{1}{L}\sum_{i=1}^{l}\sum_{j=0}^{1}((A_i == j) \odot M_i)\log(f_{\phi 2}(X_i, X_t)) \tag{3}$$

$$l_{total} = l_1 + l_2 \tag{4}$$

其中 l_1 为SED分支40分类损失，l_2 为FBSC分支前后景分类损失，x_i 和 x_t 分别表示对应于SED分支和FBSC分支的输入窗口，Y_i 和 A_i 是对应的帧级别标签，c 表示类别总数（在本研究中为40），L 是训练数据的总帧数，$f_{\phi 1}$ 和 $f_{\phi 2}$ 分别表示SED分支和FBSC分支的特征提取器，两者共享相同的4个FDC块骨干网络，$M_i \in \{0,1\}$ 表示窗口内帧的训练掩码，其中“0”表示包含截断事件的帧。符号“⊙”表示点积操作。

2.1 频率动态卷积特征提取网络

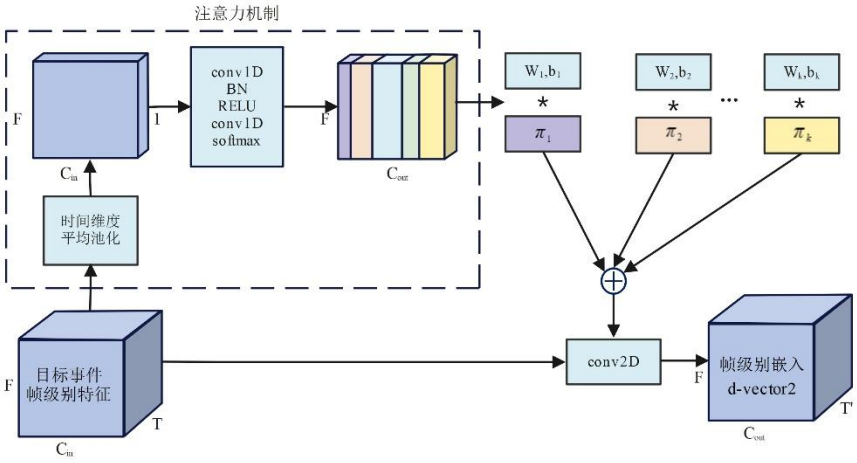


图3 频率动态卷积特征提取网络

Fig.3 Frequency Dynamic Convolutional Feature Extraction Network

在声音事件检测任务中，声音事件的时频特征通常呈现出显著的频率依赖性。在时频结构中，声音事件在时间轴上的平移仅表示事件发生时间的变化，其频率成分保持不变，因此在感知上具有等效性，而在频率轴上的平移则会改变频率成分，从而导致感知含义发生显著变化[20]。同时，传统的二维卷积在频率维度采用固定卷积核，难以适应不同事件间频率分布的差异性[20]。容易忽略微小但关键的局部特征，从而削弱模型对细微频率变动的感知能力，影响整体识别精度。

为了解决这些问题，本文在多任务检测框架中引入了FDC作为主干特征提取网络。首先预设一组可学习的基础卷积核，随后对于每个频率位置利用自适应注意力权重对这些基础核进行加权融合，动态生成契合该频率区域特征分布的卷积核。此设计既保留了时间轴上的平移等变性，又在频率轴上实现了响应自适应，使模型能够敏锐捕捉短时、低能量或重叠声源中的微小频率差异，从而显著提升了对复杂多源噪声环境下细粒度频域特征的表征能力和检测精度，FDC模块如图3所示。

首先，将目标事件帧级别特征沿时间维度进行平均池化，以聚合时序上下文信息，并沿通道维度串联两个一维卷积（1D Convolution）层，将通道数压缩为预设的基础卷积核数量。随后通过 Softmax 操作获得归

一化的频率自适应注意力权重，使其落在 [0,1] 范围内，且不同基础核的加权系数总和为1。在两层一维卷积之间，插入批量归一化（BatchNorm）与激活函数（ReLU），以此增强特征表达能力。最终，通过加权求和操作，将多个可学习的基础卷积核组合为频率自适应卷积核，用于对原始帧级别特征进行二维卷积操作，得到的输出为帧级别嵌入向量 d-vector2，计算公式如下

$$Y(t,f)=W^{T}(I)*I_{i}(t,f)+b(I)) \tag{5}$$

$$W(I)=\sum_{k=1}^{K}\pi_{k}(I)W_{k},b(I)=\sum_{k=1}^{K}\pi_{k}(I)b_{k} \tag{6}$$

$$0\le\pi_{k}(I)\le 1,\ \sum_{k=1}^{K}\pi_{k}(I)=1 \tag{7}$$

其中 t 是时间，f 是频率，$\boldsymbol{I}$ 和 $\boldsymbol{Y}$ 是一个频率动态卷积层的输入和输出，W_k 和 b_k 是第 k 个基核的权重和偏置，$\pi_k(I)$ 是第 k 个基核的频率自适应注意力权重，k 是基核的数量。

2.2 NetMamba编码器

为提升模型在处理声音事件长时序列数据时的效率，本文在FBSC分支中引入了单向 NetMamba 编码器[21]作为信息交互模块。该线性时间状态空间模型（State Space Model, SSM），兼具高效与强建模能力。与传统的多向结构相比，单向设计显著减少了冗余计算和不必要的全局信息传递。在长序列建模任务中，NetMamba 相较于 LSTM[22]及 Transformer[23]等主流模型，能有效降低计算与内存开销，缓解梯度消失，并在捕捉长期依赖关系方面表现优异，其编码器结构如图 4 所示。

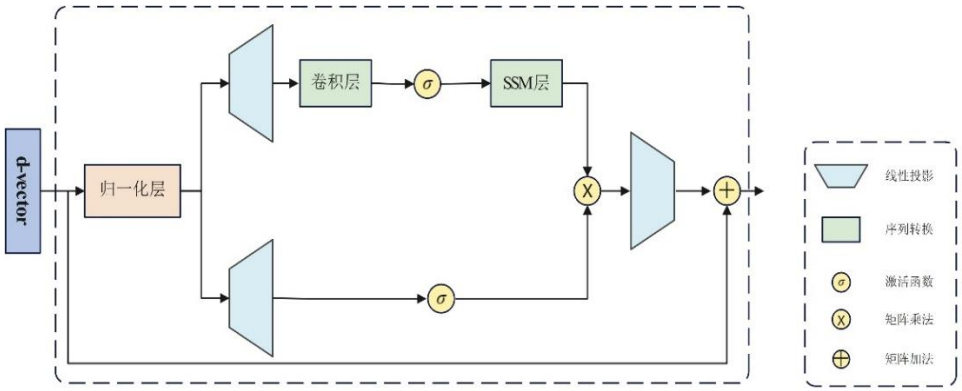


图4 NetMamba信息交互网络

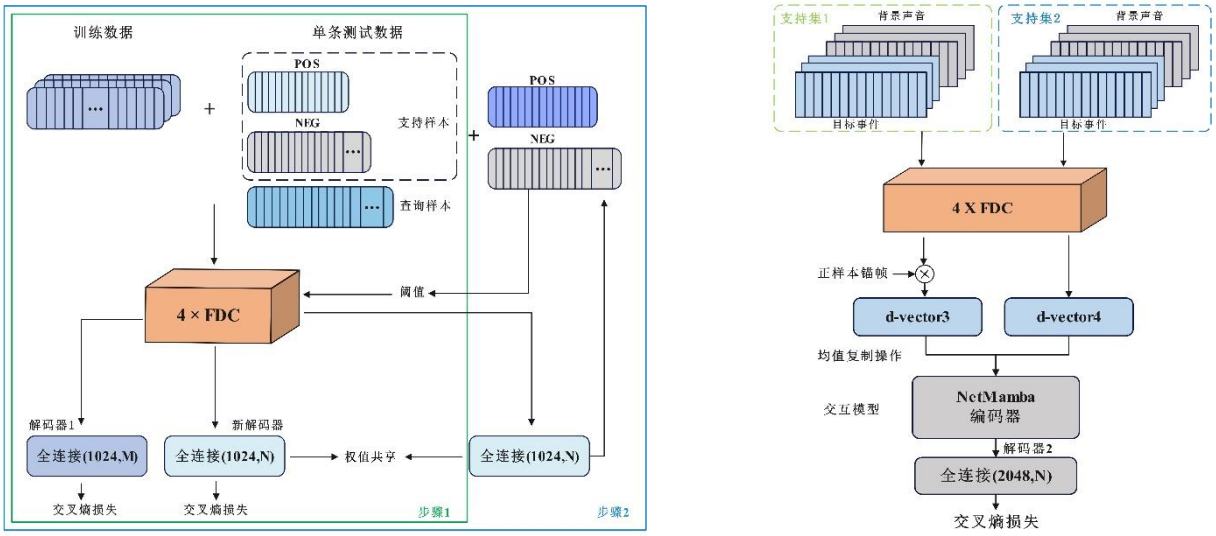
Fig.4 NetMamba Interaction Module

输入特征d-vector首先经过归一化处理（Norm），随后被送入两个并行的线性变换分支以实现特征解耦：上支路用于建模序列中的时序依赖关系，依次通过一维卷积（Conv）与状态空间模块（SSM），分别捕捉局部与长程依赖，并在中间引入激活函数以增强特征表达能力；下支路则作为门控路径，通过激活函数生成动态门控权重，用于调控上支路的输出。随后，两支路的结果通过逐元素乘法实现动态调制，并与另一线性映射路径的输出特征d-vector在加法模块中融合，最终形成融合特征表示该结构融合了状态空

基金项目：国家自然科学基金（编号：62373368，62473360）

间建模的高效性与门控机制的灵活性，在保持线性时间复杂度的同时，显著提升了模型对长时序依赖的建模能力。

2.3 多任务微调框架



（a）SED 分支　　（b）FBSC分支

图 5. 多任务微调框架

Fig.5 Multi-task Fine-tuning Framework

在模型的微调阶段，本文设计了一个双分支的多任务微调框架，如图 5 所示。该框架在保持训练阶段主干结构不变的基础上，通过少量标签样本对两个任务分支进行目标导向的参数更新，从而在新的测试环境下实现快速适应与性能提升。为增强支持集样本的时序覆盖与频率多样性，提升模型在微调过程中的泛化能力，在微调阶段引入了支持集重建与TimeFilterAug两种数据增强机制。前者通过时域重采样拓展了候选事件的上下文信息，后者通过对频谱图中若干频带施加随机增益或衰减，模拟不同声学环境下的能量分布变化。这两种增强方法应用于微调阶段的输入构造，协同提升了SED与FBSC两个分支在少样本条件下的学习表现，相关细节将在后续小节详细介绍。

在SED分支中，本文采用伪标签扩展的两步微调策略，旨在解决极少标注样本条件下正类信息不足的问题，逐步提升模型对目标事件类别的帧级别分类能力。首先，在初始阶段，从支持集中选取5个标注的正类事件片段作为正样本（POS），并在每两个POS之间间隔四个未标注片段作为负样本（NEG），构建帧级别的二元分类任务，利用交叉熵损失函数对两类进行判别训练。为增强模型的表示能力和判别能力，进一步引入辅助任务，将训练集数据与POS样本联合构建40类的多分类训练，强化特征在类别空间中的可分性。此外，考虑到训练集与支持集存在域偏移问题，本文引入了解码器迁移机制，即将训练阶段中“0类”所对应的分类器权重替换为当前任务中POS类别的归一化嵌入向量，从而在特征空间层面实现两类任务的一致性对齐。

在完成初步分类器训练后，进入第二步优化阶段。此阶段利用第一步模型对查询集进行推理，基于后验概率结果设定固定阈值，从中筛选出置信度较高的帧作为新的正样本，扩展原有支持集。随后，利用扩

展后的样本集重复前述二元分类训练。该两步过程循环迭代若干轮次，实现从有限标注样本到更大正类集合的逐步扩展与模型自适应优化。该策略显著提升了模型对目标事件类别的建模能力，尤其在处理样本稀缺的情况下仍具备良好鲁棒性, 该流程如图5（a）所示。

在FBSC分支 中，为解决支持集中仅包含有限正类帧、难以构建鲁棒类别原型的问题，采用支持集重建策略扩展有限正类帧的表示范围。通过支持集重建构建两个子集Support1和Support2，引导模型在帧级别学习前景事件的上下文特征及其与背景的区分特性。首先从支持集重建Support1 中选取包含单个POS事件与多个背景事件的音频片段，利用帧级别标签构建二值掩码，提取其中的正类帧嵌入特征，并通过均值操作构建事件目标向量d-vector3。与此同时，从Support2 中选取包含多个POS与背景事件的片段，提取帧级别嵌入特征d-vector4，输入至由训练阶段迁移而来的NetMamba编码器以获取上下文增强的特征表示。随后，d-vector3与d-vector4 通过注意力机制实现特征对齐，并送入一个全连接线性分类器，对Support2 中的每一帧进行前景/背景的二分类预测, 该流程如图5（b）所示。

2.3.1 支持集重建

由于支持集仅涵盖五种前景声音事件，模型在处理多样化声音场景时存在一定的局限性。一方面，类别数量的限制导致模型在训练过程中仅接触到有限的声音模式，难以覆盖真实环境中多样化的前景声源特征，容易出现类别偏置现象。另一方面，缺乏足够的类别间差异信息也会削弱模型对前/后景边界的判别能力，影响其在未见类别上的泛化表现。

为缓解此问题并模拟各种场景下的声音事件，本文引入了支持集重建方法。鉴于系统的帧级别特性，我们基于声音事件检测并不严重依赖非事件类NEG这一前提进行操作。通过随机改变NEG和POS的位置来丰富样本多样性。具体而言，给定支持集 $S=\{N_1,P_1,\cdots,N_5,P_5\}$ 和查询集 $Q=\{q_1,q_2,\cdots,q_i\}$ 其中 N_i，P_i 分别表示NEG和POS，q_i 表示查询集上的5秒滑动窗口，每个 q_i 输入到预训练的特征提取器 f_ϕ 中，选择与 q_i 余弦相似性低的样本以增强 N_i。此后从增强后的支持集中，随机选取包含单个 POS 事件和多个背景事件（NEG）的音频片段构建Support1。每个片段通过滑动窗口切分，确保每个窗口仅包含一个前景事件与多个背景噪声，以模拟实际应用中单一前景事件的情境。Support2 则由包含多个 POS 事件与多个背景事件的音频片段组成，旨在模拟多个前景事件同时发生的复杂场景。通过这种通过抽样和拼接的策略，支持集的样本能够覆盖更广泛的声音事件配置，从而增强模型对复杂环境的适应性。

2.3.2 时间滤波数据增强（TimeFilterAug）

在实际的推理阶段，查询音频样本往往包含远场声音、背景脉冲噪声等复杂的干扰成分[29]。为了增强声学模型对不同声学环境的适应能力，本文提出了基于随机线性滤波的数据增强策略TimeFilterAug通过随机选取log-Mel频谱图中的若干频带，并在这些频带上施加随机增益或衰减，从而模拟出不同声学环境中的各种声学能量分布。从而在不破坏原始结构信息的前提下，有效提升模型的泛化能力和环境鲁棒性。

基金项目：国家自然科学基金（编号：62373368，62473360）

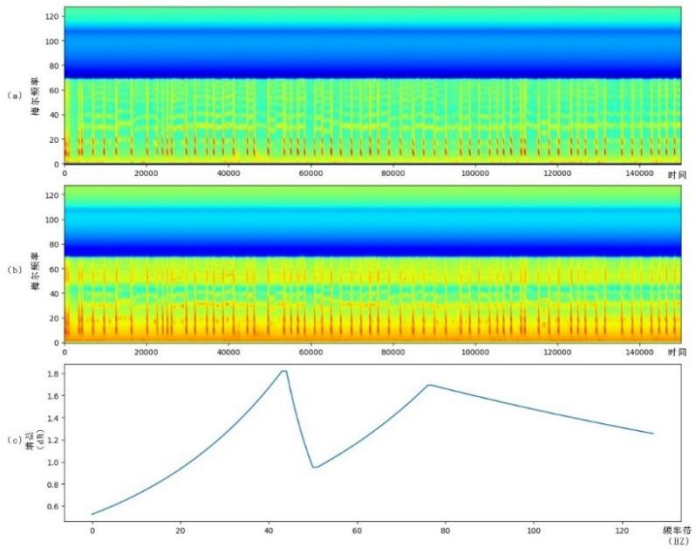


图 6. 时间滤波数据增强

Fig.6 Data augmentation using TimeFilterAug

具体而言，我们首先将输入的Log-Mel频谱图（对应约5秒的音频片段）通过滑动窗口方式随机划分为若干个时间段，记作 $T=T_1,T_2,\ldots,T_m$。然后，为每个 T_i 分配一个范围在0到1之间的随机增益权重 g_i。为了模拟频率响应的连续变化，我们进一步对每个时间段 T_i 应用一组线性变化的增益掩模，该线性掩模的函数表达式如下

$$Aug_{T_i}=T_i\cdot 10^{\beta/20} \tag{8}$$

$$\beta=l+(r-l)\cdot\alpha \tag{9}$$

$$\alpha=[g_{i-1},\ldots,g_i] \tag{10}$$

其中，Aug_{T_i} 表示增强区间，“$\cdot$”表示逐元素乘法，β 为对应的分贝增益系数，基于区间上下界 l 与 r 之间的线性插值而生成，α 表示相邻增益因子之间的插值结果。

图6展示了TimeFilterAug数据增强方法的处理效果示意图。其中，图6(a)为原始音频样本的log-Mel频谱图，图6(b)为经过TimeFilterAug频率增益曲线增强处理后的频谱图，图6(c)则展示了所施加的频率增益曲线，即模拟滤波器在不同频率带上的增益变化。可以观察到，增强后的频谱图中部分频率区域的能量有所提升或减弱，这种非均匀的频率增益模拟了在不同声学环境中，由于混响、遮蔽或远近距离引起的频率响应变化。通过这种方式TimeFilterAug能够有效提高模型在训练过程中的频域鲁棒性，促使模型学习在更宽频率范围内提取声音信息，避免过度依赖某些显著频率特征，进而提升模型在实际预测阶段应对多种背景噪声条件下的泛化能力。

3 实验结果与分析

3.1 实验设置

在训练阶段，本研究采用多任务学习策略，同步开展声音事件检测（40类）与前景/背景2分类任务。优化器选用Adam，初始学习率设定为 1e-4，并引入StepLR 学习率动态衰减策略，衰减因子Gamma与衰减步长分别为0.5和10。采用交叉熵损失函数计算损失，总训练迭代次数为100次，训练批次大小设定为32。NetMamba 编码器的卷积维度为4，状态维度为16，输入与输出维度均为2048。

在微调阶段，仅更新三个解码器以及特征提取网络的最后两层参数。FBSC分支的学习率设置为1e-4，FBSC分支的学习率则为1e-3，微调阶段的训练迭代次数为100次。为提升帧级别模型在复杂声学环境下的鲁棒性，TimefilterAug在第40次迭代后生效，将窗口分为4个时间段（即T1~T4），每个时间段的最小帧数为48，用以模拟背景噪声动态变化，提升模型在复杂声学环境下的鲁棒性与泛化能力。

3.2 DCASE 2024 任务 5 的实验结果

在DCASE 2024 Task 5评估集上，本文所提出的帧级别多任务检测系统取得了56.7%的F1-score，在全部参赛系统中排名第二。值得注意的是，该结果是在未引入FDC模块，仅采用常规CNN作为主干网络的基础上实现的，充分验证了所提多任务学习策略本身的有效性。与传统基于段级预测的方法相比，本方法通过引入帧级别预测与辅助任务建模，更精细地捕捉短时事件，显著提升了检测准确率。

表2. DCASE 2024 任务 5 上的性能比较

Table 2 PERFORMANCE COMPARISON ON DCASE2024 TASK5

方案	评估集
Hoffman et al. [24]	8.1 ± 0.3
Bordoux et al. [25]	42.0 ± 0.5
Sun et al. [26]	42.5 ± 0.5
Kao et al. [27]	46.9 ± 1
Liu et al. [28]	65.2 ± 0.3
The proposed method	56.7 ± 0.3

3.3 消融实验研究

为进一步验证所提出多任务检测框架中各子模块的独立作用及协同效应，本文设计了系统性的消融实验。实验主要从以下四个方面进行逐项剥离或替换：① 任务架构（单任务 vs 多任务）；② 时间建模结构（BiLSTM / Transformer / NetMamba）；③ 特征提取结构（传统卷积 CNN vs 频率动态卷积 FDC）；④ 数据增强策略（是否采用 TimeFilterAug）。全部实验在同一数据预处理流程和模型参数设置下进行，以确保结果可比性。实验结果见表3所示。

表3. 本文方法与其他方法对帧级别模型影响的消融研究

Table 3 The Ablation Study on the Impact of Different Methods on the Frame-Level Model

	方案				验证集
	数据增强	交互模型	主干网络	任务框架	F-score
(基线)	-	-	CNN	ST+STft	62.04

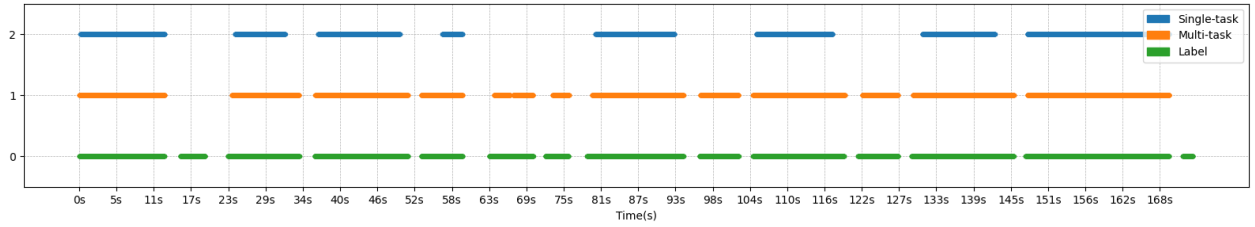
基金项目：国家自然科学基金（编号：62373368，62473360）

1	-	-	CNN	MT+STft	66.32
2	-	-	CNN	MT+MTft	67.97
3	-	Bi-LSTM	CNN	MT+MTft	68.93
4	-	Transformer	CNN	MT+MTft	70.61
5	-	NetMamba	CNN	MT+MTft	71.35
6	TimeFilterAug	NetMamba	CNN	MT+MTft	72.4
7	TimeFilterAug	NetMamba	FDC	MT+MTft	**74.49**

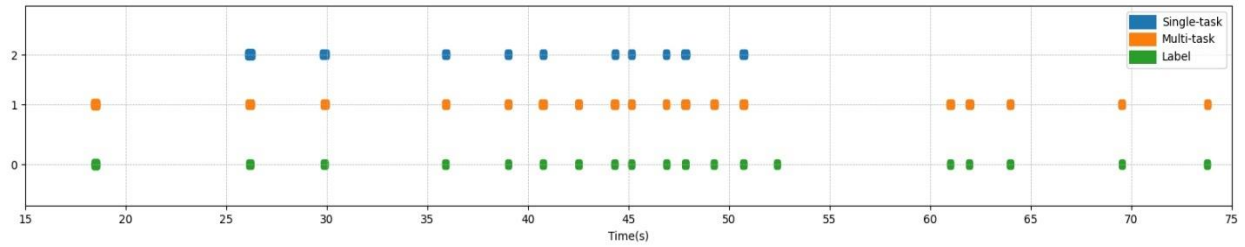
3.3.1 多任务学习的影响

在多任务学习策略方面，通过将单任务结构（ST）与多任务结构（MT）在训练阶段和微调阶段进行对比，验证了辅助任务对主任务检测性能的促进作用。结果表明，相比仅执行声音事件检测任务的单任务基线，在训练阶段引入前景/背景二分类辅助任务后，F1得分由62.04%提升至66.32%；进一步在微调阶段保持多任务结构后，性能继续提升至67.97%。这表明辅助任务通过提供帧级别结构化弱监督，引导模型更有效地区分事件与非事件帧，在一定程度上缓解了小样本训练中存在的过拟合风险与边界模糊问题，从而显著增强了模型的检测精度与泛化能力。

为进一步展示多任务学习策略对模型检测能力的增强作用，图7提供了单任务与多任务模型在处理长时与短时声音事件中的性能可视化对比。图中蓝色条带代表单任务模型的预测结果，橙色表示多任务模型的预测输出，绿色为人工标注的真实标签。从图7（a）可见，对于持续时间较长的声音事件，多任务模型展现出更为连贯的事件持续性预测，有效减少了预测过程中的中断和不稳定现象；同时，其对事件起止边界的判定更加准确，边界位置与真实标签高度对齐，反映出良好的时序建模能力。在图7（b）所示的短时声音事件示例中，尽管事件持续时间短、能量弱，多任务模型仍能准确定位事件发生的帧位置，预测响应更加集中于目标事件区域，显著提升了识别精度。此外，在多个声音事件共存的复杂场景下，多任务模型展现出更强的干扰抑制与多类事件判别能力，进一步验证了其在现实应用中的鲁棒性与实用价值。



(a)R4_cleaned recording_TEL_25-10-17.wav



(b)ME2.wav

图 7. 单任务和多任务框架在短时和长时声音事件上的性能可视化和比较。

Fig.7 The visualization and comparison of single/multi-task framework performance on short/long-duration sound events

3.3.2 时间建模结构的影响

为评估不同序列建模机制对模型时序建模能力的影响，本文在特征编码器末端分别引入了Bi-LSTM、Transformer和NetMamba三种结构进行对比。在保持主干网络与训练策略一致的条件下，实验结果显示，Bi-LSTM作为经典双向递归结构，在建模短时上下文信息方面表现稳健，将F1分数提升至68.93%；Transformer通过自注意力机制实现全局建模，使F1分数进一步提高至70.61%；而NetMamba作为新型状态空间建模结构，在保持良好建模能力的同时具有更高计算效率，最终F1分数达到71.35%。上述结果说明，增强模型的时序依赖建模能力对于提升检测性能具有关键作用，尤其是在事件发生时间短暂、背景波动显著的场景下，状态空间建模结构具备更好的效率与稳定性。

3.3.3 特征提取主干网络的影响

在特征提取模块方面，本文将常规的CNN替换为FDC，以增强模型在频率维度的自适应能力。在前置模块、时间编码结构和训练策略保持一致的条件下，实验结果表明，使用FDC作为主干特征提取器的模型在验证集上取得了 74.49% 的 F1 分数，较使用传统CNN 的对比方案提升了2.09%。FDC通过构建频率自适应卷积核，有效消除了传统卷积在频率维度上平移等方差假设所导致的感知偏差，增强了模型对频率维度的适应能力。实验结果表明，引入频率自适应机制对于提升帧级别声学特征的判别性具有显著作用。

3.3.4 TimeFilterAug 的影响

为提升模型在时变背景噪声条件下的鲁棒性，本文引入了一种基于线性扰动的时域增强策略TimeFilterAug。该策略通过将每段音频划分为多个时间区段，并引入动态扰动模拟真实背景变化，在不增加样本数量的前提下提升了训练数据的多样性与分布复杂性。实验结果表明，在引入TimeFilterAug 后，F1 分数由 71.35% 进一步提升至 72.40%，在验证集上性能提升更为显著（约 3% 至 5%）。该结果表明，通过增强模型对时间结构的适应能力，可进一步改善模型对短时瞬态事件和复杂噪声环境的响应性能，具有良好的实用推广前景。

4 结束语

针对小样本声音事件检测在样本稀缺、时序粒度不足及复杂声学环境适应性差等方面存在的挑战，本文提出了一种融合多任务学习与频率动态卷积的帧级别检测方法。该方法引入前景/背景辅助分类任务以提升事件关注能力，构建频率动态卷积模块改善特征提取器在频率维度的自适应特征提取效果，并结合基于状态空间建模的NetMamba编码器，有效提升模型在长时序上下文中的建模深度与效率。在DCASE 2024国际挑战赛Task5中，所提出方法在验证集上取得了74.49%的F1得分，在官方评估集上达到56.7%，最终排名第二。消融实验进一步验证了各子模块的独立价值与协同增益作用，表明该框架在短时瞬态事件检测、边界精准识别及复杂声学环境下的鲁棒性方面具有显著优势。本研究为低资源条件下的声音事件检测提供了一种具备推广潜力的建模范式。未来将探索跨场景泛化与时序自监督方法，进一步提升系统的适应性与实用性。

基金项目：国家自然科学基金（编号：62373368，62473360）

参考文献：

[1] Fedorishin D, Forte L, Schneider P, Setlur S, Govindaraju V. Fine-grained engine fault sound event detection using multimodal signals[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2024: 1186–1190.

[2] Sharma S, Sato K, Gautam B P. A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques[J]. Sustainability, 2023, 15(9): 7128.

[3] Fernando T, Sridharan S, Denman S, Ghaemmaghami H, Fookes C. Robust and interpretable temporal convolution network for event detection in lung sound recordings[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(7): 2898–2908.

[4] Fan X Q, Khishe M, Alqahtani A, et al. A dual adaptive semi-supervised attentional residual network framework for urban sound classification[J]. Advanced Engineering Informatics, 2024, 62: 102761.

[5] Cakır E, Parascandolo G, Heittola T, et al. Convolutional recurrent neural networks for polyphonic sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(6): 1291–1303.

[6] Wang Y, Salamon J, Bryan N J, Bello J P. Few-shot sound event detection[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2020: 81–85.

[7] Chen J M, Wang G J, Lv J C, et al. Open-set classification for signal diagnosis of machinery sensor in industrial environment[J]. IEEE Transactions on Industrial Informatics, 2022, 19(3): 2574–2584.

[8] Chou S Y, Cheng K H, Jang J S R, Yang Y H. Learning to match transient sound events using attentional similarity for few-shot sound recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2019: 26–30.

[9] Liang J H, Nolasco I, Ghani B, et al. Mind the domain gap: a systematic analysis on bioacoustic sound event detection[C]//2024 32nd European Signal Processing Conference (EUSIPCO). New York: IEEE, 2024: 1257–1261.

[10] Xia Z Q. Two-way prototypical network based on word embedding mixup for few-shot event detection[C]//2023 4th International Conference on Computer Engineering and Application (ICCEA). New York: IEEE, 2023: 46–51.

[11] Yang D, Wang H, Zou Y, et al. A mutual learning framework for few-shot sound event detection[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2022: 811–815.

[12] Shimada K, Koyama Y, Inoue A. Metric learning with background noise class for few-shot detection of rare sound events[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2020: 616–620.

[13] Banoori F, Ijaz N, Shi J L, et al. Few-shot bioacoustics event detection using transductive inference with data augmentation[J]. IEEE Sensors Letters, 2024, 8(3): 1–4.

[14] Ferroni G, Turpault N, Azcarreta J, et al. Improving sound event detection metrics: Insights from DCASE 2020[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2021: 631–635.

[15] Shi B W, Sun M, Puvvada K C, et al. Few-shot acoustic event detection via meta learning[C]//ICASSP 2020-2020 IEEE International Conference

on Acoustics, Speech and Signal Processing. New York: IEEE, 2020: 76–80.

[16] Adavanne S, Politis A, Nikunen J, Virtanen T. 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.

[17] 李云峰, 闫祖龙, 高天, 方昕, 邹亮. 基于多任务学习的语音情感识别[J]. 数据采集与处理, 2024, 39(2): 424-432.

[18] Zhang X Y, Wang S X, Du J, et al. Frame-level embedding learning for few-shot bioacoustic event detection[C]//2023 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2023: 750–755.

[19] Anderson M, Kinnunen T, Harte N. Learnable frontends that do not learn: Quantifying sensitivity to filterbank initialisation[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2023: 1–5.

[20] Xiao S C, Zhang X S, Zhang P Y. Multi-Dimensional Frequency Dynamic Convolution with Confident Mean Teacher for Sound Event Detection[R]. ICASSP 2023, New York: Tech. Rep., 2023.

[21] Wang T Z, Xie X H, Wang W D, et al. Netmamba: Efficient network traffic classification via pre-training unidirectional mamba[C]//2024 IEEE 32nd International Conference on Network Protocols (ICNP). New York: IEEE, 2024: 1–11.

[22] Greff K, Srivastava R K, Koutník J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(10): 2222–2232.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[24] Hoffman B, Robinson D. Toward in-context bioacoustic sound event detection[R]. DCASE2024 Challenge, New York: Tech. Rep., 2024.

[25] Bordoux V. ADAPTABLE INPUT LENGTH USING MODEL TRAINED ON WAVEFORM[R]. DCASE2024 Challenge, New York: Tech. Rep., 2024.

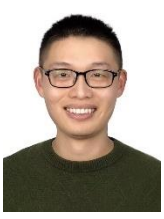
[26] Sun M K, Zhang H J, Qian K, et al. LIF-PROTONET: Prototypical network with leaky integrate-and-fire neuron and squeeze-and-excitation blocks for bioacoustic event detection[R]. DCASE2024 Challenge, New York: Tech. Rep., 2024.

[27] Kao S L, Liu Y W. Cosine similarity based few-shot bioacoustic event detection with automatic frequency range identification in Mel-Spectrograms[R]. DCASE2024 Challenge, New York: Tech. Rep., 2024.

[28] Liu W, Liu H Y. Few-shot Bioacoustic Event Detection at the DCASE 2024 Challenge[R]. DCASE2024 Challenge, New York: Tech. Rep., 2024.

[29] Nam H, Kim S H, Park Y H. Filteraugment: An acoustic environmental data augmentation method[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2022: 4308–4312.

作者简介：



方昕(1988-)，男，博士，工程师，研究方向：语音识别、说话人识别，E-mail：xinfang@iflytek.com

赵鹏源(2000-)，男，硕士，研究方向：深度学习、声音事件检测。

张雨涛(2003-)，男，硕士，研究方向：计算机视觉、多模态数字水印。

基金项目：国家自然科学基金（编号：62373368，62473360）

严根伟(1997-)，男，硕士，研究方向：声音事件检测、小样本学习

闫祖龙(2000-)，男，硕士，研究方向：机器学习、语音情感识别。

邹亮(1987-)，通信作者，男，博士，副教授，研究方向：深度学习、信号处理，E-mail：liangzou@cumt.edu.cn