融合预训练表征与伪强标签的声音事件检测方法

关晓菡　罗伟杰 李嘉锋 蔡希昌
（北方工业大学 人工智能与计算机学院 北京市 100144）

中图分类号：TN912.3;TP18

摘要

声音事件检测（Sound Event Detection, SED）性能提升的关键瓶颈在于对精确时间标注强标签数据的高度依赖。然而，真实场景中采集的音频数据通常呈现强标签、弱标签、无标签及软标签等多源异构监督形式，标注粒度与监督强度差异显著，导致模型难以统一建模并充分挖掘细粒度时序信息。针对上述问题，本文从统一优化角度出发，提出一种融合预训练声学表征与伪强标签学习的半监督声音事件检测方法，将多源异构监督信息建模为显式监督、一致性约束与伪标签监督的联合优化过程。

在模型结构上，引入 Transformer 预训练声学表征与 CNN-RNN 检测网络进行多尺度特征融合，以提升模型的全局语义建模能力与局部时频分辨能力；在训练策略上，设计渐进式两阶段微调机制，通过“特征适配—联合优化”的优化路径，实现预训练表征与帧级检测任务之间的稳定对齐；在数据利用层面，提出多模型一致性驱动的伪强标签推断机制，通过预测融合与时间结构重建，将弱标签与无标签数据转化为高质量帧级监督信号，并进一步用于轻量化模型的训练，实现基于数据驱动的知识迁移。

在 DCASE 2024 Task 4 数据集（DESED、MAESTRO Real 及 AudioSet Strong）上的实验结果表明，所提出方法能够有效提升模型性能，其中 CRNN-ATST 模型的 PSDS1 和 mpAUC 分别达到 0.531 和 0.740，较基线提升 14.8%和 9.3%。在引入伪强标签后，轻量化 LightSED 模型在参数量压缩 88.6%的情况下，总体性能仅下降 3.46%，且 PSDS2 等指标与大模型基本持平。结果表明，本文方法在提升异构监督数据利用效率的同时，兼顾检测性能与模型复杂度，为强标签稀缺及资源受限条件下的声音事件检测提供了一种有效解决方案。

关键词：声音事件检测；预训练声学表征；异构监督学习；伪强标签；半监督学习；轻量化模型

北京市自然科学基金（项目编号：IS23053）

A method for sound event detection that integrates pre-trained representations with pseudo-strong labels

Guan Xiaohan, Luo Weijie, Li Jiafeng, Cai Xichang

(North China University of Technology, Artificial Intelligence and Computer Science, Beijing 100144, China)

Abstract

The key bottleneck hindering performance improvements in Sound Event Detection (SED) lies in its heavy reliance on strongly labelled data with precise temporal annotations. However, audio data collected in real-world scenarios typically presents a multi-source, heterogeneous supervision format comprising strongly labelled, weakly labelled, unlabelled, and soft-labelled data. Significant variations in annotation granularity and supervision strength make it difficult for models to establish a unified representation and fully exploit fine-grained temporal information. To address these issues, this paper proposes a semi-supervised sound event detection method that integrates pre-trained acoustic representations with pseudo-strong labelling learning, from a unified optimisation perspective. It models multi-source heterogeneous supervision as a joint optimisation process involving explicit supervision, consistency constraints, and pseudo-labelling supervision.

In terms of model architecture, Transformer-based pre-trained acoustic representations are integrated with a CNN-RNN detection network to achieve multi-scale feature fusion, thereby enhancing the model's global semantic modelling capabilities and local spatio-temporal resolution. Regarding training strategies, a progressive two-stage fine-tuning mechanism is designed; through an 'feature adaptation–joint optimisation' pathway, stable alignment is achieved between the pre-trained representations and the frame-level detection task. In terms of data utilisation, we propose a pseudo-strong label inference mechanism driven by multi-model consistency. Through prediction fusion and temporal structure reconstruction, this mechanism converts weakly labelled and unlabelled data into high-quality frame-level supervision signals, which are subsequently utilised for training lightweight models, thereby achieving data-driven knowledge transfer.

Experimental results on the DCASE 2024 Task 4 datasets (DESED, MAESTRO Real and AudioSet Strong) demonstrate that the proposed method effectively enhances model performance. Specifically, the CRNN-ATST model achieves PSDS1 and mpAUC scores of 0.531 and 0.740 respectively, representing improvements of 14.8% and 9.3% over the baseline. Following the introduction of pseudo-strong labels, the lightweight LightSED model exhibits a mere 3.46% decline in overall performance despite an 88.6% reduction in parameters, whilst metrics such as PSDS2 remain on par with those of large models. The results demonstrate that our method balances detection performance with model complexity whilst enhancing the utilisation efficiency of heterogeneous supervised data, thereby providing an effective solution for sound event detection under conditions of scarce strong labels and resource constraints.

Keywords: Sound Event Detection; Pre-trained acoustic representations; heterogeneous supervised learning; pseudo-strong labels; semi-supervised learning; lightweight models

引言：

声音事件检测[1]旨在从连续音频流中识别特定声音事件的类别并精确定位其起止时间[2]，是智能家居、环境监测、自动驾驶辅助系统及安防监控等领域的重要基础技术[3][4]。随着物联网和边缘计算的快速发展，SED 系统对实时性、鲁棒性和泛化能力的要求日益提高[5][6]。然而，制约 SED 性能进一步提升的核心瓶颈之一[7]，是对精确时间标注的强标签数据的高度依赖[8]。强标签的获取成本高昂[10]，尤其在复杂真实场景中，事件起止时间的准确标注难度极大[11]，导致实际采集的音频数据往往同时包含强标签、弱标签、无标签以及软标签等多种异构监督形式[12]。不同标注粒度与监督强度的显著差异，不仅增加了模型优化的难度，也对声学特征的鲁棒性与泛化能力提出了更高要求[13]。

传统基于卷积循环神经网络（CRNN）[14]的 SED 方法虽能有效捕捉局部时频结构[15]，但其特征学习高度依赖有限的任务特定数据。在强标签样本不足或存在分布偏移的情况下[16]，模型难以习得稳定且具有迁移能力的通用声学表征，从而限制了整体检测性能[17][18]。近年来，大规模音频分类数据集（如 AudioSet[19]）的出现以及 Transformer[20]网络架构的迭代，推动了预训练声学模型的快速发展。BEATs[29]、ATST-Frame[22]、CED[21]等基于 Transformer 的预训练模型在大规模音频数据上通过自监督或有监督方式学习到丰富的语义特征[26][27]，并在下游音频分类任务中取得领先性能[28]。鉴于声音事件分类与 SED 在声学语义建模上的高度相关性[25]，将此类预训练模型迁移至 SED 任务，有望缓解强标签稀缺问题并提升模型泛化能力[30]。

然而，现有的预训练模型多采用片段级判别学习，其表征粒度、时间分辨率及训练目标与 SED 帧级检测任务仍存在明显差异，直接影响迁移效果[10]。因此，有必要系统评估不同预训练声学表征在 SED 中的适用性，并设计针对性的微调与训练策略，以充分发挥其在异构监督场景下的潜力[31]。同时，真实场景中采集的多源异构音频数据（合成强标签、真实弱标签、无标签及软标签）蕴含着大量潜在时序信息，如何高效挖掘这些数据并转化为可靠的帧级监督信号，成为当前 SED 领域亟待解决的关键问题。

针对上述问题，提出一种融合预训练声学表征与伪强标签学习的半监督声音事件检测方法。以统一优化框架为基础，将多源异构监督信息建模为显式监督、一致性约束与伪标签监督的协同学习过程，通过结构设计与训练策略优化，实现对不同标注粒度数据的有效利用与时序信息的充分挖掘。本文的主要贡献如下：

1. 提出一种融合预训练声学表征的多尺度声音事件检测框架，通过结合 Transformer 预训练模型与 CNN-RNN 结构，实现全局语义信息与局部时频特征的协同建模，从而提升模型在复杂声学环境下的表征能力与泛化性能；

2. 提出一种渐进式表征—任务对齐的两阶段微调策略，通过“特征适配—联合优化”的分阶段训练机制，缓解预训练模型与帧级检测任务之间的分布差异，并结合一致性约束提升无标签数据的利用效率；

3. 提出一种多模型一致性驱动的伪强标签学习方法，通过预测融合与时间结构重建，将弱标签与无标签数据转化为高质量帧级监督信号，实现从隐式一致性学习到显式监督增强的转化；

4. 构建一种基于伪标签驱动的轻量化模型训练方法，在显著降低模型参数

规模的同时保持较高检测性能，为资源受限场景下的声音事件检测提供有效解决方案。

1. 方法

1.1 整体框架

针对声音事件检测中强标签稀缺及多源异构监督信息难以统一建模的问题，本文从统一优化角度出发，将不同标注粒度的数据表示为显式监督、一致性约束与伪标签监督的联合学习问题。整体优化目标可表示为：

$$L = L_{BCE} + \lambda_1 L_{MSE} + \lambda_2 L_{Pseudo} \quad (1)$$

其中，L_{BCE}表示基于强标签与弱标签的监督损失，L_{MSE}表示一致性约束损失，L_{Pseudo}表示由伪强标签引入的显式监督项,在该统一框架下，本文方法由三个相互协同的关键模块构成具体如图 1 所示：

(1)特征提取模块：通过融合 Transformer 预训练模型与 CNN 分支，实现全局语义信息与局部时频结构的协同建模；

(2)渐进式时序对齐模块：通过两阶段微调策略，实现预训练特征分布与帧级检测任务之间的稳定对齐；

(3)伪强标签时间建模模块：通过多模型一致性推断与时间结构重建，将弱标签与无标签数据转化为帧级显式监督信号。

在结构上，模型采用双分支特征提取方式：CNN 分支用于提取局部时频模式，预训练分支用于建模长时依赖关系。两类特征经时间对齐后进行融合，并输入序列建模模块进行帧级预测。该框架通过统一优化机制实现不同监督信号的协同利用，从而提升模型对复杂声学事件的时序建模能力与泛化性能。

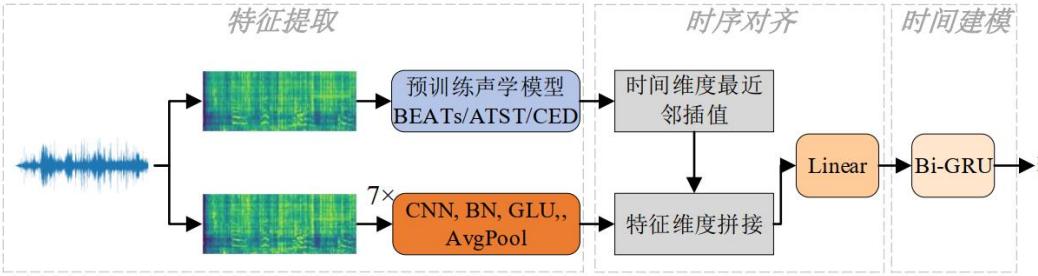

图 1 与预训练声学表征结合的模型框架图

1.2 预训练声学表征结合的声音事件检测模型

为提升声音事件检测模型在强标签数据不足条件下的表征能力，本文在传统 CRNN 结构基础上，引入基于 Transformer 的预训练声学模型作为辅助特征提取分支，构建多尺度特征融合的检测框架。该模型结构来源于声音事件分类领域的预训练模型，并针对帧级检测任务进行了适配设计。

1.2.1 预训练声学模型

本文选取三类具有代表性的预训练声学模型作为特征提取器，包括 BEATs、ATST-Frame 以及 CED。这三类模型均在大规模音频数据（AudioSet）上进行预训练，并在声音事件分类任务中表现出较强的性能与泛化能力。

在结构上，上述模型均基于 Transformer 编码器构建，其核心由多层自注意力模块堆叠组成。输入音频首先转换为对数梅尔谱图表示，并进一步划分为固定大小的 patch 序列，通过线性投影映射为高维嵌入向量，同时加入位置编码以保

留时频位置信息。随后，嵌入序列输入多层 Transformer 编码器，通过多头自注意力机制建模长时程依赖关系，并结合前馈网络进行非线性变换，最终输出高层声学表征。尽管三类模型在结构上基本一致，但其预训练策略存在显著差异：

BEATs 模型采用自监督掩码建模与离散声学标记联合优化策略。在训练过程中，通过声学标记器生成离散标签，并利用掩码预测任务优化音频编码器；随后通过知识蒸馏反向优化标记器，从而形成迭代优化过程。该方法能够逐步提升声学表示的语义一致性与判别能力。

ATST-Frame 模型基于教师—学生框架，通过构建同一音频的不同视图，引入时间维掩码机制，在帧级层面施加一致性约束。该方法通过仅对被掩码帧计算损失，使模型能够学习更细粒度的时序表示，从而更适用于帧级检测任务。

CED 模型采用一致性集成蒸馏策略，通过融合多个教师模型的预测结果作为软标签，并在多种数据增强条件下约束学生模型输出一致性，从而提升模型的鲁棒性。该方法能够有效缓解单一教师模型带来的偏差问题。

上述三类预训练模型在结构统一的前提下，通过不同训练策略学习到具有差异性的声学表示，为后续分析其在声音事件检测任务中的迁移能力提供基础。

1.2.2 融合检测模型结构

在预训练声学表征的基础上，本文构建了融合 CNN 与 Transformer 特征的声音事件检测模型。

对输入音频信号进行对数梅尔谱图变换，得到特征表示：

$$\mathrm{X} \in R^{T \times F} \tag{2}$$

其中，T 为时间帧数，F 为频率维度。该表示作为模型两个分支的输入。

CNN 分支用于提取局部时频结构信息。该分支由多层卷积层、批归一化及激活函数组成，并通过池化操作逐步降低时间与频率分辨率，从而获得具有一定感受野的局部特征表示：

$$H_{cnn} = f_{cnn}(X) \tag{3}$$

该特征能够有效刻画声音事件的局部频谱模式与短时动态变化。预训练分支利用 Transformer 模型提取高层语义特征。输入谱图经过 patch 划分与线性映射后，输入编码器进行特征提取，得到H_{pt}，该特征在时间维度上具有较强的上下文建模能力，能够捕捉长时依赖关系。由于 CNN 分支与预训练分支在时间分辨率上存在差异，需要对预训练特征进行对齐。为实现跨分辨率特征对齐，采用插值操作将预训练特征映射至与 CNN 分支一致的时间尺度，该操作在保证计算效率的同时，实现不同时间分辨率特征的统一表示。

$$H_{pt} = f_{pt}(X), \qquad \widetilde{H}_{pt} = \mathrm{Interp}\left(H_{pt}\right) \tag{4}$$

对齐后的预训练特征与 CNN 特征在通道维度进行拼接得到H，融合特征同时包含局部时频信息与全局语义信息。为保证后续模块计算效率，对融合特征进行线性映射与维度压缩：

$$H = Concat\left(H_{cnn}, \widetilde{H}_{pt}\right), \qquad H' = WH + b \tag{5}$$

融合特征输入双向循环神经网络（Bi-GRU）进行时序建模，最终通过线性层输出帧级预测结果，其中，Y 表示每一时间帧上各类别的预测概率。

$$S = f_{rnn}(H'), \qquad Y = \sigma(W_O S + b_O) \tag{6}$$

1.3 渐进式微调策略

在引入预训练声学模型后，其表征分布与帧级声音事件检测任务之间存在显著差异。预训练模型通常基于片段级判别任务进行优化，其学习到的特征主要面向全局语义表达，而 SED 任务则依赖于细粒度的时间定位能力。这种“表征粒度不匹配”问题，使得直接进行端到端联合微调往往导致训练不稳定，甚至破坏预训练模型中已学习到的通用声学表示。

从优化角度来看，该问题可归结为预训练特征分布与任务特定分布之间的对齐问题。同时，在半监督学习框架下[35]，无标签数据通过一致性约束参与训练，其梯度具有较高不确定性，当预训练模型与随机初始化的检测网络同时更新时，不稳定梯度可能进一步干扰特征表示的迁移过程。

基于上述分析，本文提出一种渐进式表征—任务对齐的两阶段微调策略，通过分阶段优化实现从“特征适配”到“任务对齐”的平滑过渡，从而提升训练稳定性与模型泛化能力。

1.3.1 特征适配阶段

在第一阶段训练中，冻结预训练声学模型参数，仅优化任务相关的 CRNN 检测网络。此时，预训练模型作为固定的特征提取器，其输出定义了一个稳定的声学表示空间，模型优化过程可表示为：

$$\min_{\theta_{task}} L\left(f_{\theta_{task}}\left(\tilde{H}_{pt}\right)\right) \tag{7}$$

该阶段，冻结预训练声学模型的全部参数，仅对 CRNN 检测网络进行训练。此时，预训练模型作为固定的高质量特征提取器，为后续网络提供稳定的声学表示。

该阶段的核心目标是使检测网络逐步适应预训练特征的分布特性，在稳定的特征空间中学习声音事件的判别规则与基本时序结构。由于预训练特征具有较强的语义表达能力，CRNN 网络可以在较低的训练难度下建立起初步的检测能力。

在损失函数设计上，该阶段同时利用有标签数据与无标签数据进行训练。对于强标签与弱标签数据，采用二元交叉熵损失作为监督信号；对于无标签数据，则基于 Mean Teacher 框架构建一致性约束[33]，通过最小化学生模型与教师模型预测结果之间的差异，实现对无标签数据的利用。

在损失函数设计上，采用监督损失与一致性约束联合优化：

$$L_{stage1} = L_{BCE} + \lambda L_{MSE} \tag{8}$$

其中，L_{BCE}表示基于强标签与弱标签的监督损失，L_{MSE}表示学生模型与教师模型之间的一致性损失，λ为权重系数。

需要指出的是，在该阶段中一致性损失的权重较低，其主要作用是对模型输出施加平滑约束，避免在有限标注数据下发生过拟合。通过该阶段训练，检测网络能够在稳定特征空间中建立基本的时序建模能力。

1.3.2 联合微调与一致性强化

在完成第一阶段训练后，模型已在预训练特征空间中建立初步的判别能力。在此基础上，进入第二阶段训练，通过解冻预训练模型参数，实现端到端联合优化：

$$\min_{\theta_{task},\ \theta_{pt}} L\left(f_{\theta_{task}, \theta_{pt}}(\mathrm{X})\right) \tag{9}$$

与第一阶段不同，首先加载第一阶段训练得到的模型参数，并解冻预训练声学模型，使其与 CRNN 网络共同参与训练。该阶段的优化目标由“特征适配”转向“表征—任务对齐”，即通过联合更新预训练模型与检测网络参数，使特征表示逐步向帧级检测任务所需的时间分辨率与判别结构调整。

为了进一步提升无标签数据的利用效率，本文在第二阶段引入插值一致性训练（Interpolation Consistency Training, ICT）[34]，并与 Mean Teacher 框架结合，对模型施加更强的一致性约束。具体而言，通过对无标签样本进行输入空间插值，并约束模型在插值样本上的预测与原始样本预测的线性组合保持一致，从而提升模型在输入空间的平滑性与泛化能力。对无标签数据施加更强的一致性约束：

$$L_{stage2} = L_{BCE} + \lambda_1 L_{MSE} + \lambda_2 L_{ICT} \tag{10}$$

其中，L_{ICT}表示插值一致性损失，λ_1与λ_2分别为对应损失项的权重系数。

相比第一阶段，该阶段显著提高一致性损失的权重，使无标签数据在优化过程中占据更重要作用，从而引导模型学习潜在的时间结构信息，实现对有限强标签数据的有效补充。

1.4 伪强标签辅助的半监督声音事件检测方法

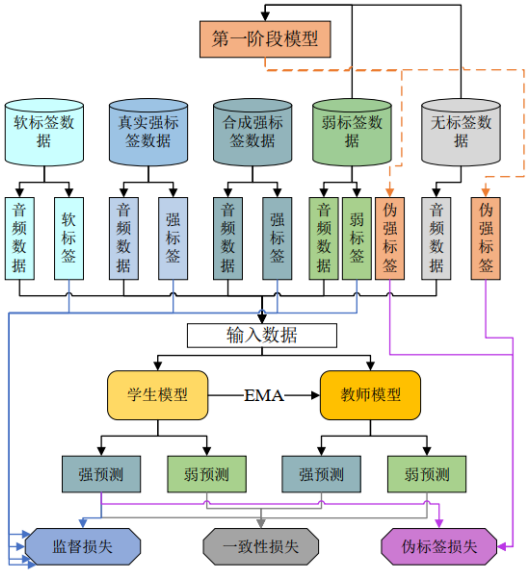

图 2 伪强标签辅助训练流程

在 1.3 节所述两阶段微调策略中，模型主要通过一致性约束实现对无标签数据的利用。然而，该类方法本质上属于隐式正则化，仅约束模型在扰动下预测的一致性，缺乏对声音事件时间边界的直接监督，因而在帧级检测任务中对细粒度时序结构的建模能力仍然有限。

为进一步挖掘无标签数据中的潜在时序信息，本文提出一种不确定性感知的伪强标签学习机制（Uncertainty-aware Pseudo-Strong Labeling, UPSL）。该方法通过构建“预测—筛选—重构—再训练”的闭环优化流程，将无标签数据转化为高质量帧级监督信号，从而实现从隐式一致性约束向显式时序监督的转化。整体流程如图 2 所示。

1.4.1 伪强标签生成方法

在初始模型训练完成后，利用该模型对弱标签与无标签数据进行前向推理，获得对应的帧级事件概率序列。由于单一模型在预测过程中可能受到数据分布偏

差或训练噪声的影响，直接使用其输出构建伪标签容易引入误差。因此，本文采用多模型集成策略对预测结果进行融合，以提升伪标签的稳定性与可靠性。

设第 k 个模型对某一音频样本的预测结果为$P_k(t,c)$，则通过对多个模型输出进行平均融合，可以得到更为平滑且鲁棒的预测结果：

$$P(t,c)=\frac{1}{K}\sum_{k=1}^{K}P_k(t,c) \tag{11}$$

其中，t 表示时间帧，c 表示事件类别，K 为参与集成的模型数量。该方法能够有效降低单模型预测偏差，使生成的伪标签更加接近真实标签分布。

在此基础上，为进一步评估预测结果的可靠性，引入基于模型分歧的不确定性估计机制：

$$U(t,c)=\frac{1}{K}\sum_{k=1}^{K}\left(P_k(t,c)-P(t,c)\right)^2 \tag{12}$$

其中，$U(t,c)$表征模型预测的不确定性水平。当不同模型预测差异较大时，不确定性升高，说明该位置的伪标签可信度较低。

类似于 FixMatch[36]中基于高置信度预测进行伪标签筛选的方法，本文进一步结合模型不确定性构建联合筛选机制。在伪标签生成过程中，首先通过模型预测概率$P(t,c)$表示帧级类别置信度，再根据类别自适应阈值 τ进行二值化筛选：

$$\hat{Y}(t,c)=\begin{cases}1 & P(t,c)>\tau\\ 0 & P(t,c)\leq\tau\end{cases} \tag{13}$$

其中，τ为类别自适应阈值，用于缓解类别不平衡问题。具体实现时，可根据每类概率分布的分位数或均值动态设定阈值，从而保证低频类别也能选出高置信帧作为伪标签。经过筛选的二值帧序列再进行时间连续合并，生成包含起止时间信息的事件级伪标签。相比传统中值滤波平滑方法，该策略更精确地保留事件边界，减少时间偏移误差。

相比传统基于中值滤波的平滑方法，该策略能够更准确地保留事件边界信息，避免过度平滑带来的时间偏移问题，从而提高伪标签的时间精度与一致性。

1.4.2 伪强标签联合训练机制

在生成伪强标签后，将其与原有强标签数据共同用于模型训练，从而构建多源监督融合的优化框架。在该阶段中，训练数据同时包含真实帧级标签与模型生成的伪标签，使模型能够在更大规模数据上学习时间定位信息。

为统一不同来源的监督信号，本文在原有损失函数基础上引入伪标签监督项，整体优化目标可表示为：

$$L=L_{BCE}+\lambda_1 L_{MSE}+\lambda_2 L_{Pseudo} \tag{14}$$

$$L_{Pseudo}=\sum_{t,c}w(t,c)\cdot BCE\left(Y(t,c),\hat{Y}(t,c)\right) \tag{15}$$

$$w(t,c)=\frac{P(t,c)}{1+U(t,c)} \tag{16}$$

其中L_{Pseudo}表示伪强标签监督损失，权重$w(t,c)$根据预测置信度与不确定性动态调整。与仅依赖一致性约束的方法相比，该设计使模型在训练过程中更加关

注高置信度、低不确定性的伪标签，从而有效降低噪声标签对模型优化的干扰。

通过引入显式伪标签监督，原有“仅约束输出一致性”的半监督学习过程被扩展为“显式标签增强”的监督学习过程，从而显著提升模型对时间边界的建模能力。

1.4.3 伪标签驱动的轻量化模型训练

在实际部署场景中，声音事件检测模型通常需要运行于边缘设备或资源受限平台，因此除了检测性能外，模型参数规模与推理效率同样重要。尽管融合预训练声学表征的 CRNN-ATST 模型能够获得较高检测性能，但其参数量约为 100M，计算开销较大，不利于实际部署。为验证本文所提出伪强标签学习机制在轻量化场景下的有效性，进一步设计轻量化检测模型 LightSED。

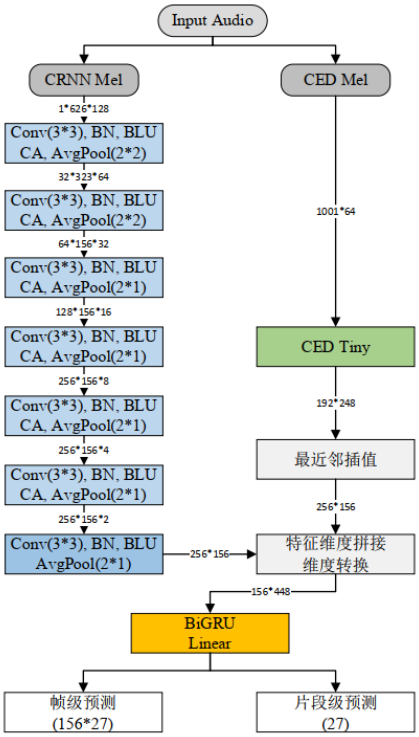

图 3 LightSED 网络结构

LightSED 整体结构如图 3 所示，仍采用“特征提取—时序建模—帧级分类”的检测框架，但针对特征提取模块进行了轻量化设计。具体而言，采用深度较浅的卷积网络替代大规模预训练 Transformer 编码器，并减少卷积通道数与网络层数，以降低模型参数量和计算复杂度。同时保留双向 GRU 作为时序建模模块，以保证模型对声音事件时间依赖关系的建模能力。最终通过全连接层输出帧级事件概率。

与 CRNN-ATST 相比，LightSED 不再依赖大规模 Transformer 预训练表征，而是通过紧凑卷积结构直接学习局部时频特征，从而显著降低模型复杂度。由于模型容量受限，LightSED 在仅使用真实标签训练时性能存在一定下降。为弥补其表征能力不足，本文利用性能较优的 CRNN-ATST 模型生成伪强标签，并将其作为额外监督信号参与训练。通过引入高质量帧级伪标签，轻量化模型能够学习到更准确的时间边界信息，实现从大模型到小模型的数据驱动知识迁移。

实验统计结果表明，LightSED 参数量约为 10.1M，仅为 CRNN-ATST 模型

的 10.1%，参数压缩率达到 89.9%。

2. 实验设置

2.1 数据集

本章实验基于 DCASE 2024 Task 4 的多源异构数据构建训练与评测环境，包括 DESED、MAESTRO Real 和 AudioSet Strong[37]。DESED 提供强标签、弱标签及无标签数据，用于基础监督与一致性学习；MAESTRO Real 提供概率形式软标签，以刻画标注不确定性，并在评估时通过阈值转换为硬标签；AudioSet Strong 提供精确时间边界信息，作为补充强标签数据。针对类别分布差异，本文对 MAESTRO Real 筛选 11 类主要事件包括鸟鸣、汽车声、人声、脚步声、儿童说话声、风吹声、刹车声、大型车辆声、餐具声、地铁接近声和地铁离开声。并对 AudioSet Strong 建立类别映射关系，统一至任务定义类别空间。三类数据在监督粒度与强度上形成互补，为模型在半监督与伪强标签框架下的训练提供多层次监督信息，从而提升声音事件检测的时序建模能力与泛化性能。

2.2 参数设置

在实验过程中，所有数据集统一重采样至 16 kHz，并将音频片段长度固定为 10 s。对于 CNN 特征提取分支，输入音频**按照本文实验设置**转换为 128 个梅尔滤波器组的梅尔谱图，时间帧数为 626。对于预训练模型分支，则保持各模型原始预训练阶段的特征提取配置，以保证声学表征的一致性。本文采用分阶段学习率策略：在冻结预训练特征提取模块参数时，仅训练新加入的任务相关层，学习率设置为 1×10^{-3}；在解冻全部网络参数进行端到端微调阶段，将学习率降低至 1×10^{-4}，以保证参数更新的稳定性并避免破坏预训练权重的已学习表征。

2.3 评价指标

在评价指标方面，本文采用 PSDS1、PSDS2、mpAUC 及 CB-F1 对模型性能进行综合评估。其中，PSDS1 与 mpAUC 作为主要指标，分别衡量模型的时间定位能力与片段级判别能力；PSDS2 与 CB-F1 作为辅助指标，用于分析模型在不同时间容差条件下的检测性能。

针对不同数据集的标注形式，采用差异化评估策略：对于具有帧级强标签的 DESED 数据集，采用 PSDS 指标评估模型的事件级检测性能，重点反映时间边界定位能力；而对于仅提供片段级软标签的 MAESTRO Real 数据集，由于缺乏精确时间戳，无法进行事件级匹配，因此采用 mpAUC 指标，从片段层面评估模型对声音事件的区分能力。

3 实验结果与分析

本章围绕预训练声学表征引入、微调策略优化及伪强标签辅助学习三个方面开展对比实验。在统一训练设置下，逐步引入各模块以评估不同策略对声音事件检测性能的影响。

首先，为验证预训练声学表征的迁移有效性，以 CRNN 作为基线模型，分别引入 ATST-Frame、BEATs 和 CED-Base 作为特征提取分支，构建 ATST-CRNN、BEATs-CRNN 和 CED-CRNN 模型。在该实验中，冻结预训练模型参数，仅使用其提取的声学表征作为输入，以隔离预训练特征本身对 SED 性能的贡献。

表 1 预训练声学表征对声音事件检测性能的影响

模型	PSDS1	mpAUC	PSDS2	CB-F1
CRNN	0.383	0.647	0.641	0.452
BEATs-CRNN	**0.497**	0.683	**0.788**	**0.589**
ATST-CRNN	0.495	0.706	0.768	0.539
CED-CRNN	0.431	**0.708**	0.735	0.520

表 1 给出了不同模型的性能对比结果。相较于基线 CRNN，引入预训练声学表征后，各项指标均显著提升，表明预训练特征能够有效增强模型的表征能力。其中，BEATs-CRNN 在 PSDS1、PSDS2 及 CB-F1 指标上取得最优性能，表现出较强的事件检测与时间定位能力；ATST-CRNN 在各项指标上表现均衡，尤其在 PSDS1 等时间定位相关指标上具有竞争优势；CED-CRNN 在 mpAUC 指标上表现最佳，但在 PSDS1 等帧级检测指标上的提升相对有限。

进一步分析表明，尽管三种预训练模型在结构与训练数据上基本一致，其性能差异主要来源于预训练策略与下游任务的匹配程度。BEATs 通过离散声学标记与掩码建模联合优化，兼顾语义建模与时序信息表达；ATST-Frame 通过帧级一致性约束强化时间建模能力，使其更适用于帧级检测任务；而 CED 侧重全局语义蒸馏，虽在片段级判别任务中表现较好，但对细粒度时间边界的建模能力相对不足。

总体而言，预训练策略与 SED 任务在时间分辨率与监督形式上的匹配程度，是影响迁移效果的关键因素。

3.1 消融实验

3.1.1 微调策略对检测性能的影响

在完成预训练声学表征迁移能力的对比实验后，本节进一步分析微调策略对模型性能的影响。考虑到 BEATs 与 ATST-Frame 在前述实验中表现最优，且显著优于 CED-Base，本文选取二者作为预训练特征提取器开展消融研究。

表 2 给出了不同微调策略下的模型性能对比结果。可以观察到，两阶段微调策略在两种预训练模型上均显著优于直接联合微调。以 BEATs 为例，PSDS1 由 0.382 提升至 0.506；ATST-Frame 亦呈现一致趋势，其 PSDS1 由 0.265 提升至 0.507，表明两阶段策略能够显著提升模型的时间定位能力。

针对 ATST-Frame，我们注意到直接联合微调性能显著下降（0.265，相较于单独 ATST-CRNN 微调的 0.495）。这一异常下降的主要原因在于训练过程中使用了无标签数据，会对梯度产生干扰。具体而言：在联合微调中，无标签数据的伪监督信号尚不稳定，容易与已有的 ATST 预训练表征发生梯度冲突，从而破坏帧级事件判别能力。其次，ATST 预训练强调全局语义特征，而联合微调伪标签主要作用于局部帧级事件，目标不一致加剧了训练不稳定性。 直接联合微调缺乏渐进式约束，ATST 权重在前期更新过于激烈，导致原有表征被破坏，性能异常下降。

两阶段微调策略通过“先冻结、后联合”的优化过程，先在稳定特征空间中训练检测网络，再逐步解冻预训练模型实现端到端优化，有效缓解了上述梯度干扰问题。实验结果显示，该策略在提升 PSDS1 与 CB-F1 等强调时间边界建模能力的指标上效果显著，而对 PSDS2 与 mpAUC 等偏重分类性能或宽松匹配条件的指标影响相对有限。这表明，两阶段微调主要提升了细粒度时序建模能力，而

非单纯增强分类判别能力。

综上，两阶段微调不仅能显著提升时间定位性能，也能稳定预训练特征迁移至下游 SED 任务的效果，避免了直接联合微调可能导致的性能异常下降。

表 2 不同微调策略对声音事件检测性能的影响

模型	微调策略	PSDS1	mpAUC	PSDS2	CB-F1	可训练参数量
CRNN-BEATs	直接联合微调	0.382	0.722	0.634	0.288	104M
	二阶段微调	0.506	0.710	**0.800**	**0.595**	104M
CRNN-ATST	直接联合微调	0.265	**0.767**	0.587	0.304	100M
	二阶段微调	**0.507**	0.713	0.786	0.557	100M

3.1.2 伪强标签辅助训练策略对检测性能的影响

在验证两阶段微调策略有效性后，进一步引入伪强标签辅助学习，以增强对弱标签与无标签数据中潜在时序信息的利用能力。考虑到 ATST-Frame 在关键指标上表现最优，本文以其作为主模型，并结合 BEATs 进行集成推理，生成帧级伪强标签。

如表 3 所示，引入伪强标签后模型性能获得稳定提升。其中，PSDS1 由 0.507 提升至 0.531，mpAUC 由 0.713 提升至 0.740，CB-F1 由 0.557 提升至 0.598。该结果表明，伪强标签通过提供额外的帧级显式监督信号，有效弥补了强标签数据不足对时间建模能力的限制，从而提升整体检测性能。

表 3 伪强标签辅助训练策略对声音事件检测性能的影响

模型	伪强标签	PSDS1	mpAUC	PSDS2	CB-F1
CRNN-ATST	×	0.507	0.713	0.786	0.557
CRNN-ATST	√	0.531	0.740	0.804	0.598

进一步地，将该策略应用于轻量化模型 LightSED，结果如表 4 所示。在未引入伪强标签时，LightSED 性能显著低于 CRNN-ATST；而在引入伪强标签后，其性能获得明显提升，PSDS1 由 0.416 提升至 0.507，PSDS2 由 0.726 提升至 0.798，CB-F1 提升至 0.569。值得注意的是，在参数量仅为 10.1M 的情况下，LightSED 在 PSDS2 指标上已接近 CRNN-ATST（100M）的性能水平。任务中，相比单纯增加模型容量，引入更多可靠的帧级监督信息往往能够带来更为显著的性能提升。

表 4 轻量化检测模型在伪强标签辅助下的性能比较

模型	参数量	伪标签	PSDS1	mpAUC	PSDS2	CB-F1
LightSED	10.1M	×	0.416	0.723	0.726	0.523
LightSED	10.1M	√	0.507	0.720	0.798	0.569
CRNN-ATST	100M	√	0.531	0.740	0.804	0.598

上述结果表明，伪强标签能够在数据层面提供有效的知识迁移路径，使轻量化模型在有限容量下仍能学习到较为准确的时间边界信息。相比单纯增加模型规模，引入高质量帧级监督信号在提升模型性能方面更为有效，尤其在多源异构数据场景下，其优势更加明显。

3.2 与现有方法对比

为进一步验证本文方法的有效性与先进性，本文将所提出方法与近年来声音事件检测领域具有代表性的公开方法进行对比，比较结果如表 5 所示。对比方法包括 DCASE2024 Task4 中具有代表性的系统，如 Zhang_BUPT、LEE_KT 与

Nam_KAIST 等。所有方法均在相同评价指标下进行比较，其中 PSDS1 用于衡量模型在严格时间边界条件下的检测能力，mpAUC 则用于评估整体事件分类性能。

从表 5 可以看出，本文方法在 PSDS1 与 mpAUC 等指标上均取得一定的优势。其中，PSDS1 达到 0.531，相较于官方基线（0.475）提升约 11.7%，表明本文方法能够更有效地建模声音事件的时间边界信息；同时，mpAUC 达到 0.740，相较于官方基线（0.646）提升约 14.6%，说明模型在片段级事件判别能力方面同样具有较强性能。

表 5 与现有代表性方法在 DCASE 2024 Task 4 数据集上的性能对比结果的性能比较

方法	PSDS1	mpAUC
Baseline[38]	0.475	0.646
Zhang_BUPT[39]	0.528	0.702
LEE_KT[40]	0.449	0.683
Nam_KAIST[41]	0.526	0.772
本文方法	0.531	0.740

4 结束语

面向声音事件检测中强标签稀缺及多源异构监督难以统一建模的问题，本文提出一种融合预训练声学表征与伪强标签的半监督学习方法。该方法以 Transformer 预训练模型为表征基础，结合 CNN-RNN 结构实现多尺度时序建模，并通过两阶段微调策略缓解预训练任务与帧级检测目标之间的分布不匹配问题。在此基础上，构建伪强标签生成与利用机制，通过多模型集成与时间边界建模，将弱标签与无标签数据转化为显式帧级监督信号，从而提升模型对事件时序边界的刻画能力。同时，进一步将该机制扩展至轻量化模型训练，实现性能与模型复杂度之间的有效权衡。实验结果表明，所提方法在 DCASE 2024 任务数据上显著提升 PSDS 与 mpAUC 等指标，并在参数受限条件下仍保持较强性能，验证了其在实际应用场景中的有效性与可扩展性。

This work was supported in part by the Beijing Natural Science Foundation under Grant IS23053.

参考文献：

[1] Mesaros A, Heittola T, Virtanen T, et al. Sound event detection: A tutorial[J]. IEEE Signal Processing Magazine, 2021, 38(5): 67-83.

[2] Yin H, Chen J, Bai J, et al. Multi-granularity acoustic information fusion for sound event detection[J]. Signal Processing, 2025, 227: 109691.

[3] Yan J, Song Y, Guo W, et al. A region based attention method for weakly supervised sound event detection and classification[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 755-759.

[4] Liu Peng, Zhang Zhenya, Wang Ping,等.基于深度学习的声音事件检测综述[J].安庆师范大学学报(自然科学版),2025,31(03):33-39.DOI:10.13757/j.cnki.cn34-1328/n.2025.03.006.

[5] Ohta T, Bando Y, Imoto K, et al. A sequential audio spectrogram transformer for real-time sound event detection[C]//2024 32nd European Signal Processing Conference (EUSIPCO). Lyon, France: IEEE, 2024: 101-105..

[6] Zheng Y, Yao L. Research on Traffic Sound Event Detection Based on Multi-Scale Feature Fusion[J]. Applied Sciences, 2026, 16(5): 2359.

[7] Li Haitao.基于深度学习的弱标注多声音事件检测研究事件检测研究[D].青岛科技大学,2022.DOI:10.27264/d.cnki.gqdhc.2022.000052.

[8] Xiao S, Zhang X, Zhang P, et al. Semi-supervised sound event detection with dynamic convolution and confidence-aware mean teacher[J]. Digital Signal Processing, 2025, 156: 104794.

[9] Tang Yiqi.基于半监督学习的声音事件检测[D].南京邮电大学,2024.DOI:10.27251/d.cnki.gnjdc.2024.000630.

[10] Schmid F, Primus P, Morocutti T, et al. Multi-iteration multi-stage fine-tuning of transformers for sound event detection with heterogeneous datasets[EB/OL]. https://arxiv.org/abs/2407.12997, 2024-07-18.

[11] Pankajakshan A, Bear H L, Benetos E. Polyphonic sound event and sound activity detection: A multi-task approach[C]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, USA: IEEE, 2019: 323-327.

[12] Xu L, Wang L, Bi S, et al. Semi-supervised sound event detection with pre-trained model[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[13] Cui H, Song L, Li L, et al. Revisiting SSL for sound event detection: complementary fusion and adaptive post-processing[J]. Journal of King Saud University Computer and Information Sciences, 2025, 37(6): 160.

[14] Shao Yanbing, Zhang Zhengwei, Zhang Xuming.基于改进 CRNN 网络的城市声音事件检测算法[C]//中国自动化学会过程控制专业委员会,中国自动化学会.第 36 届中国过程控制会议论文集（上）.浙江大学控制科学与工程学院

智能感知与检测技术研究所;,2025:1176.DOI:10.26914/c.cnkihy.2025.113235.

[15] Shao N, Li X, Li X. Fine-tune the pretrained ATST model for sound event detection[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 911-915.

[16] Chen Haojie, Yang Rui, Pan Shanliang.基于 MAML 改进的少样本声音事件检测算法[J].数据采集与处理,2025,40(3):741-753.DOI:10.16337/j.1004-9037.2025.03.014.

[17] Zheng Y, Zhang R, Atito S, et al. ASiT-CRNN: A method for sound event detection with fine-tuning of self-supervised pre-trained ASiT-based model[J]. Digital Signal Processing, 2025, 160: 105055.

[18] Chen Pengfei, Xia Xiuyu.基于迁移学习卷积记忆网络的多声音事件检测[J].数据采集与处理,2025,40(3):730-740.DOI:10.16337/j.1004-9037.2025.03.013.

[19] Gemmeke J F, Ellis D P W, Freedman D, et al. Audio set: An ontology and human-labeled dataset for audio events[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 776-780.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[21] Dinkel H, Wang Y, Yan Z, et al. CED: Consistent Ensemble Distillation for Audio Tagging[EB/OL]. (2023-08-23). https://arxiv.org/abs/2308.11957.

[22] Gong Y, Chung Y A, Glass J. AST: Audio spectrogram transformer[EB/OL]. https://arxiv.org/abs/2104.01778, 2021-04-05.

[23] Li K, Song Y, Dai L R, et al. AST-SED: An effective sound event detection method based on audio spectrogram transformer[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[24] Li K, Song Y, McLoughlin I, et al. Fine-tuning audio spectrogram transformer with task-aware adapters for sound event detection[C]//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2023). Dublin, Ireland: International Speech Communication Association (ISCA), 2023: 291-295.

[25] Zhao Ming, Chen Rui.一种基于特征融合的声音事件检测方法[J].数据采集与处理,2025,40(06):1556-1567.DOI:10.16337/j.1004-9037.2025.06.014.

[26] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449-12460.

[27] Serizel R, Turpault N, Shah A, et al. Sound event detection in synthetic domestic environments[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 86-90.

[28] Hsu W N, Shi B. u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality[J]. Advances in Neural Information Processing Systems, 2022, 35: 21157-21170.

[29] Chen S, Wu Y, Wang C, et al. BEATs: Audio pre-training with acoustic tokenizers

[EB/OL]. https://arxiv.org/abs/2212.09058, 2022-12-18.

[30] Xiao Y, Yin H, Bai J, et al. MixStyle based domain generalization for sound event detection with heterogeneous training data[EB/OL]. https://arxiv.org/abs/2407.03654, 2024-07-04.

[31] Cai P, Song Y, Li K, et al. MAT-SED: A masked audio transformer with masked-reconstruction based pre-training for sound event detection[EB/OL]. https://arxiv.org/abs/2408.08673, 2024-08-15.

[32] Liu Yafen, Zheng Yifeng, Jiang Lingyi,等.深度半监督学习中伪标签方法综述[J].计算机科学与探索,2022,16(06):1279-1290.

[33] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. Advances in neural information processing systems, 2017, 30.

[34] Verma V, Kawaguchi K, Lamb A, et al. Interpolation consistency training for semi-supervised learning[J]. Neural Networks, 2022, 145: 90-106.

[35] Lee D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]//Workshop on challenges in representation learning, ICML. 2013, 3(2): 896.

[36] Sohn K, Berthelot D, Carlini N, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence[J]. Advances in neural information processing systems, 2020, 33: 596-608.

[37] Turpault N, Serizel R, Shah A P, et al. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis[C]//Workshop on Detection and Classification of Acoustic Scenes and Events. 2019.

[38] Cornell S, Ebbers J, Douwes C, et al. DCASE 2024 task 4: Sound event detection with heterogeneous data and missing labels[EB/OL]. https://arxiv.org/abs/2406.08056, 2024-06-12.

[39] Yue H, Wang Z, Mu D, et al. Local and global features fusion for sound event detection with heterogeneous training dataset and potentially missing labels[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE 2024). Tokyo, Japan, 2024.

[40] Lee Y, Jung J H. Technical report on Lee submission: sound event detection using conformer and ATST framework for DCASE Challenge 2024 Task 4[R]. DCASE Challenge 2024 Technical Report, 2024.

[41] Nam H, Min D, Choi S, et al. Self-training and ensembling frequency-dependent networks with coarse prediction pooling and sound event bounding boxes[EB/OL]. https://arxiv.org/abs/2406.15725, 2024-06-22.

作者信息：

姓名：关晓菡
照片：
出生年月:

性别：
职称：
研究方向：
邮箱：

姓名：罗伟杰
照片：
出生年月:
性别：
职称：
研究方向：

姓名：李嘉锋
照片：
出生年月:
性别：
职称：
研究方向：

通信作者：
姓名：蔡希昌
照片：
出生年月:
性别：
职称：
研究方向：
邮箱：

英文长摘要

Addressing the challenges in sound event detection tasks—namely the high cost of acquiring strongly labelled data, the underutilisation of heterogeneous supervised data, and the limited transferability of pre-trained models—this paper investigates a semi-supervised sound event detection method based on pre-trained representations and pseudo-strong labelling. By integrating pre-trained acoustic models with semi-supervised learning strategies, the method enhances the model's ability to detect sound event categories and temporal boundaries in complex acoustic scenarios, whilst exploring effective methods for transferring high-performance models to lightweight ones.

This paper constructs a semi-supervised sound event detection framework that integrates pre-trained representations with pseudo-ground-truth learning. Firstly, pre-trained models such as ATST, BEATs and CED are employed to extract high-level acoustic features, which are then integrated with a CRNN detection network. Secondly, a two-stage fine-tuning strategy is designed to enhance the model's transferability through a combination of freezing and joint fine-tuning. Subsequently, frame-level pseudo-strong labels are generated using a teacher model, and the quality of these pseudo-labels is improved through multi-model ensemble, uncertainty-aware weighting, and temporal structure reconstruction strategies. Finally, a lightweight model, LightSED, is designed to validate the knowledge transfer effectiveness of pseudo-strong label learning in lightweight scenarios.

Experiments were conducted using the DCASE Task 4 public dataset. The results indicate that pre-trained representations can significantly improve model detection performance, with ATST-CRNN achieving the best detection results. The two-stage fine-tuning further enhances model stability and detection capabilities. Following the introduction of pseudo-strong label learning, the model's ability to recognise event temporal boundaries is further improved. For lightweight models, pseudo-strong labels can effectively compensate for performance losses caused by insufficient model capacity, achieving detection results close to those of large models despite a lower parameter scale.

The semi-supervised sound event detection method proposed in this paper effectively integrates heterogeneous supervised data with pre-trained acoustic feature information. The two-stage fine-tuning strategy enhances the transfer performance of pre-trained models, whilst the pseudo-strong labelling learning mechanism improves the model's ability to perceive event boundaries and facilitates knowledge transfer from high-performance models to lightweight models. Experimental results validate the effectiveness of the proposed method in terms of detection performance and model deployment, providing new insights for research on semi-supervised sound event detection.

Research Highlights

(1) A semi-supervised sound event detection framework is proposed that integrates pre-trained acoustic representations with pseudo-strong label learning, enabling the unified utilisation of heterogeneous supervised data.

(2) A two-stage fine-tuning strategy is designed to effectively enhance the transferability and training stability of pre-trained models, such as ATST, in sound event detection tasks.

(3) A pseudo-strong labelling generation method is proposed, combining multi-model ensemble, uncertainty-aware weighting and temporal structure reconstruction, to improve the quality of pseudo-labels and the effectiveness of supervision.

(4) A lightweight sound event detection model, LightSED, is constructed to achieve data-driven knowledge transfer from high-performance teacher models to lightweight student models.

Keywords: Sound Event Detection; Pre-trained acoustic representations; heterogeneous supervised learning; pseudo-strong labels; semi-supervised learning; lightweight models