计算机科学
COMPUTER SCIENCE

基于元学习的半监督声音事件检测方法

沈雅馨，高利剑，毛启容

引用本文

沈雅馨，高利剑，毛启容. 基于元学习的半监督声音事件检测方法[J]. 计算机科学，2025，52(3)：222-230.

SHEN Yaxin, GAO Lijian , MAO Qirong. Semi-supervised Sound Event Detection Based on Meta Learning [J]. Computer Science, 2025, 52(3): 222-230.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100191

基于元学习的半监督声音事件检测方法

沈雅馨[1]　高利剑[1]　毛启容[1,2]
1 江苏大学计算机科学与通信工程学院　江苏 镇江 212013
2 江苏省大数据泛在感知与智能农业应用工程研究中心　江苏 镇江 212013
(2212108045@stmail.ujs.edu.cn)

摘　要　现有的半监督声音事件检测方法直接使用强标签合成样本、弱标签真实样本和无标签真实样本进行训练,以缓解标签样本量不足的问题。然而,合成和真实数据域之间存在不可避免的分布差异,这种差异会干扰模型梯度优化方向,从而限制模型的泛化能力。针对这一问题,基于元学习(Meta Learning)提出了一种新颖的半监督声音事件检测学习范式 MMT(Meta Mean Teacher)。具体来说,对于每个训练批次的数据,将其分为由合成样本组成的元训练集和由真实样本组成的元测试集;将模型在元训练集上计算的元梯度作为元测试梯度更新的指导,使模型感知并学习到更具泛化性的知识。在 DCASE2021 任务 4 数据集的测试集上进行对比实验,结果表明,相较于官方基线,所提出的学习范式 MMT 在 F1,PSDS1 和 PSDS2 指标上分别提升了 8.9%,6.6%和 1.1%;相较于当前的先进方法,所提出的学习范式 MMT 同样表现出了显著的性能优势。

关键词:声音事件检测;元学习;一致性正则化;半监督学习;深度学习
中图分类号　TP391

Semi-supervised Sound Event Detection Based on Meta Learning

SHEN Yaxin[1],GAO Lijian[1] and MAO Qirong[1,2]
1 College of Computer Science and Communication Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China
2 Jiangsu Province Big Data Ubiquitous Perception and Intelligent Agriculture Application Engineering Research Center,Zhenjiang,Jiangsu 212013,China

Abstract　Existing semi-supervised sound event detection methods directly utilize strongly labeled synthetic samples,weakly labeled real samples,and unlabeled real samples for training to alleviate the issue of insufficient labeled samples. However,there is an inevitable distribution gap between synthetic and real domains,which can interfere with the direction of model gradient optimization,thereby restricting generalization ability of these models. To address this challenge,a novel semi-supervised sound event detection learning paradigm,meta mean teacher(MMT),is proposed based on meta-learning. Specifically,for each batch of training data,it is divided into a meta-training set consisting of synthetic samples and a meta-test set consisting of real samples. The meta-gradient calculated on the meta-training set serves as guidance for updating the meta-test gradient,allowing the model to perceive and learn more generalized knowledge. Experimental results on the DCASE2021 Task 4 dataset show that,compared to the official baseline,the proposed learning paradigm MMT has a relative improvement of 8.9%,6.6%,and 1.1% in the F1,PSDS1,and PSDS2 metrics,respectively. Compared to the current state-of-the-art methods in the field,the proposed learning paradigm MMT still demonstrates a significant performance advantage.

Keywords　Sound event detection,Meta learning,Consistency regularization,Semi-supervised learning,Deep learning

1　引言

声音事件检测任务(Sound Event Detection,SED)旨在识别出给定音频信号中出现的目标声音事件类别,诸如狗叫声、警报声、汽车声等,并准确定位这些类别对应的声音事件实例在该音频中发生的起始时间和终止时间,如图 1 所示。声音事件检测具有广阔的应用前景,包括道路交通监测[1]、公共安全和安保[2],以及生态环境监测[3]等。

自 2013 年以来,国际权威组织 IEEE 举办了声学场景和事件检测及分类比赛(Challenge on Detection and Classifica-

到稿日期:2024-01-29　返修日期:2024-05-23
基金项目:国家自然科学基金(62176106);江苏省研究生科研与实践创新计划项目(KYCX22_3668);江苏大学应急管理学院专项科研项目(KY-A-01)
This work was supported by the National Natural Science Foundation of China(62176106),Postgraduate Research & Practice Innovation Program of Jiangsu Province(KYCX22_3668) and Special Scientific Research Project of School of Emergency Management,Jiangsu University(KY-A-01).
通信作者:毛启容(mao_qr@ujs.edu.cn)

tion of Acoustic Scenes and Events，DCASE)，将声音事件检测任务纳入赛道，极大地推动了该领域的发展。自此，声音事件检测成为声学领域的一大研究热点，涌现出了众多解决方案。研究者们从使用传统隐马尔可夫模型再到使用深度学习模型[4]进行事件的分类及时间定位，显著提升了性能，但深度模型的训练往往依赖于大量可靠的标注数据。对于声音事件检测任务而言，需要检测事件的时间定位信息，即在对一段音频进行人工数据标注时，需要逐帧判断其中包含的事件类型，以提供精细的监督信息。然而，这种逐帧的强标签时间戳信息获取成本高、耗时长[5]。因此，自 2018 年以来，依托平均教师(Mean Teacher，MT)框架[6]的半监督声音事件检测方法(Semi-supervised Sound Event Detection，SSED)应运而生。该类方法使用无标签和仅有分类标签而无定位时间戳信息的弱标签真实音频样本，并引入通过合成前景事件和背景场景来获得准确时间戳信息的强标签合成数据，缓解对强标签真实数据的依赖[7]。

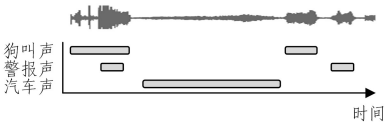


图 1 声音事件检测任务

Fig. 1 Sound event detection task

在这种半监督学习方式下，现有方法[8-12]通常将合成数据作为真实数据的强标签样本补充进行训练，期望利用合成数据提供的强标签信息，约束网络在缺乏足量的具有有效定位信息的真实数据的情况下实现更好的事件定位和分类效果。然而，合成数据和真实数据之间存在不可避免的分布差异，这种域不匹配的差异会限制模型从合成域泛化到真实域的性能[13]。为了解决这一问题，Yang 等[14]采用了基于多任务学习(Multi-task Learning)的方式，通过对抗思想，利用梯度反转层提高模型的域判别能力，从而增强模型的检测性能。但这种方式并不能有效地利用无标签的真实数据。因此，Zheng 等[15]提出了一种基于相互平均教学(Mutual Mean Teaching)的半监督声音事件检测方法，将 SSED 任务分解成两个子任务：真实数据和合成数据的域适应任务以及真实数据的半监督学习任务。前者用于缩小真实数据和合成数据之间的域差异，后者则充分利用无标签真实数据来学习真实域的数据分布。最后，两个子任务之间的输出相互约束，分别为各自提供不同的数据互补视图。然而，这种方式需要同时训练 4 个检测模型，参数量较大且未充分考虑大量强标签合成数据的知识。因此，设计一个能够考虑域差异问题且充分利用大量合成数据的声音事件检测模型变得至关重要。

在机器学习领域中，通过元学习[16]的思想训练模型可以赋予其更强的泛化能力，使其能够更快速地适应和学习新任务或新领域，从而实现域适应和域泛化的目标。这一思想的有效性已经在医学图像分割[17-18]、噪声标签学习[19-21]以及推荐系统[22-23]等任务上得到充分的验证。因此，针对上述问题，本文提出了一种新颖的基于元学习的半监督声音事件检测学习范式(Meta Mean Teacher，MMT)，通过优化模型参数更新方向的方式，实现合成域到真实域的泛化，提升模型检测性能。具体而言，对于每个训练批次的数据，将其分为由强标签合成样本组成的元训练集以及由弱标签和无标签真实样本组成的元测试集。在每个训练步中，首先在元训练阶段利用元训练集训练学生和教师模型，使用元梯度来更新学生和教师模型参数，使其能够快速适应声音事件检测任务；其次在元测试阶段使用元测试集训练学生模型，测试其在真实数据上的泛化能力，并基于一致性正则技术进一步校准学生模型梯度更新方向，通过指数移动平均更新教师模型。本文的贡献如下：

1)提出一种基于元学习的半监督声音事件检测学习范式，将模型在合成域上计算的元梯度作为真实域上梯度更新的指导，从而使模型感知并学习更具泛化性的知识，提升了半监督声音事件检测模型的泛化性能。

2)在 DCASE2021 任务 4 数据集上将所提方法与官方基线和当前先进方法进行对比，实验结果表明，基于元学习的半监督学习范式表现出显著的性能优势，充分证明了其有效性和泛化性。

本文第 2 章介绍了半监督声音事件检测和元学习的相关工作；第 3 章详细阐述了基于元学习的半监督声音事件检测学习范式 MMT；第 4 章介绍了实验设置；第 5 章给出了实验结果及性能评价，证明了所提方法的有效性；第 6 章对所提方法进行了讨论；最后总结全文并对下一步工作进行了展望。

2 相关工作

2.1 半监督声音事件检测框架

早期，研究者基于对抗训练的思想来提升声音事件检测模型的域适应性能。然而，这种方法并未充分利用易获得的无标签数据[24]。当前主流的半监督声音事件检测方法通过半监督领域的经典框架——平均教师(Mean Teacher，MT)进行构建。MT 框架结构如图 2 所示，其主要目的是更好地利用大量强标签合成数据进行监督，利用无标签真实数据进行标签泛化，将学生模型结合监督学习和一致性假设，即对于同一输入样本，迫使学生模型和教师模型的输出一致，从而有效利用无标签数据来提高模型的泛化能力。

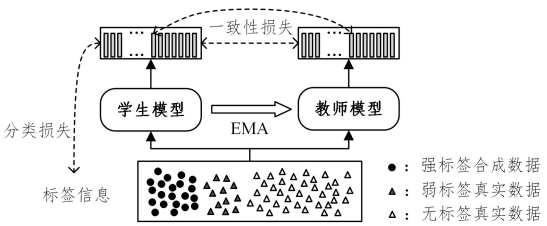


图 2 基于平均教师框架的半监督声音事件检测

Fig. 2 SSED based on MT framework

教师模型通过计算学生模型的指数移动平均(Exponential Moving Average，EMA)来更新网络参数，其目的在于寻找更稳定的梯度更新方向，减少梯度噪声，使教师模型对于输入数据更为稳定，从而更好地引导学生模型进行泛化训练。例如，2023 年提出的 Joint-Former[25]集成 MT 和掩蔽重构建模，通过联合正则技术取得了很好的检测性能提升。

2.2 元学习

元学习(Meta Learning)是机器学习领域的一项重要

研究方法，其核心目标是让模型学习如何学习（Learn-to-Learn），这个过程旨在使模型在面对不同的任务时能够快速适应，从而提高模型的泛化能力。元学习可以从很多角度去理解[26]，其中一种角度是从双层优化（Bilevel Optimization）的观点出发，揭示元学习实质上是在一个外层优化问题中嵌套一个内层优化问题，在这两个问题的优化过程中存在主从不对称性（Leader-follower Asymmetry），即内层优化过程受制于外层的学习策略。在机器学习领域，元学习有着广泛的应用，其算法可以大致分为基于优化、基于度量、基于模型3类[27]。其中，基于优化的元学习算法具有很强的泛化性，经典的MAML算法[28]（Model-Agnostic Meta-Learning）就归于该类。该算法的核心思想是在元训练集上计算模型的元梯度，然后通过在元测试集上计算参数的二阶导数矩阵（Hessian）来更新模型，从而提升模型的泛化性能。

3 基于元学习的半监督声音事件检测

3.1 半监督声音事件检测问题描述

半监督声音事件检测方法的训练数据集$\{D_s, D_w, D_u\}$包含3个子数据集：强标签合成样本集$D_s=\{(x_s^1, y_s^1), \cdots, (x_s^i, y_s^i), \cdots, (x_s^N, y_s^N)\}$、弱标签真实样本集$D_w=\{(x_w^1, y_w^1), \cdots, (x_w^i, y_w^i), \cdots, (x_w^M, y_w^M)\}$和无标签真实样本集$D_u=\{x_u^1, \cdots, x_u^i, \cdots, x_u^K\}$。其中，$x^i \in \mathbb{R}^{T\times F}$表示从第$i$个音频样本中提取的FBank频谱特征，其特征维度为F，频谱特征序列长度为T；N, M, K分别代表3个子数据集中样本的数量；$y_w^i \in \mathbb{R}^C$是维度为C的真实数据弱标签，表示第i段真实音频中出现的事件类别信息（不包含时间信息），C为声音事件类别个数；$y_s^i \in \mathbb{R}^{T\times C}$是合成数据强标签，包含第$i$段合成音频中出现的事件类别以及对应的起止时间戳信息。对于一段音频序列x，将其输入到声音事件检测器f中，得到预测结果$f_\theta(x) \in \mathbb{R}^{T\times C}$，其中$\theta$表示模型参数。在强标签中，每个样本既包含事件分类信息，也包含定位时间戳，因此可以进行帧级别（Frame-level）损失计算；而在弱标签中，每个样本只包含事件分类信息，仅能进行句级别（Clip-level）的事件分类损失计算。

对于有标签样本x_s和x_w，通过二值交叉熵（Binary Cross Entropy，BCE）计算监督分类损失$\mathcal{L}_{cls}$，其计算式如式(1)所示：

$$\mathcal{L}_{cls}=y\log f_\theta(x)+(1-y)\log(1-f_\theta(x)) \tag{1}$$

其中，$\mathcal{L}_{cls}$表示学生模型输出$f_\theta(x)$与标签y计算的交叉熵损失函数，$(x, y)\in\{D_s, D_w\}$。接着，通过一致性正则化技术，在整个数据集D上使用均方误差计算无监督对比损失$\mathcal{L}_{con}$，如式(2)所示：

$$\mathcal{L}_{con}=\frac{1}{N}\sum_{i=1}^{N}(f_{\theta_s}(x)-f_{\theta_t}(x))^2 \tag{2}$$

其中，$\mathcal{L}_{con}$表示学生模型输出$f_{\theta_s}(x)$和教师模型输出$f_{\theta_t}(x)$之间计算的一致性正则损失函数，$x\in D$，N表示样本总数；θ_s和θ_t分别表示学生模型和教师模型的参数，这两个模型具有相同的网络结构。

最后，学生模型根据式(3)进行更新。

$$\mathcal{L}_{total}=\mathcal{L}_{cls}+\lambda(t)\mathcal{L}_{con} \tag{3}$$

其中，$\mathcal{L}_{total}$表示最终用于更新学生模型的损失函数，$\lambda(t)$表示随训练迭代次数增加而缓慢增长的调和权重。教师模型则通过对学生模型的指数移动平均（EMA）进行更新，如式(4)所示：

$$\theta_t \leftarrow \alpha\theta_{t-1}+(1-\alpha)\theta_s \tag{4}$$

其中，α为EMA算法中定义的衰减率。

3.2 元梯度优化的半监督声音事件检测

利用元学习提高模型的泛化性能的核心在于明确地将模型置于具有域差异的环境中，使其感知这种域差异并学习到更有泛化性的知识[29]。其关键在于将训练数据划分为不重叠的元训练集和元测试集，以模拟域差异，通过对模型在元训练集上计算的元梯度进行优化，从而减小该模型在元测试集上的测试误差。具体而言，首先从构成训练集$D_{train}=\{D_1, \cdots, D_N\}$的$N$个可见域中随机采样数据，构建包含多个任务的任务集作为元训练集$D_{meta\text{-}train}$（$D_{meta\text{-}train}<D_{train}$），在元训练集上训练使得模型能够从多个任务中学习通用特征或模式；其次构建元测试集$D_{meta\text{-}test}=D_{train}-D_{meta\text{-}train}$，用于测试经过元训练更新后的模型在元测试任务上的泛化能力，使模型能够快速地适应新任务或领域[30]。在半监督声音事件检测任务中，简化了任务集的构造过程，在同一训练步内，将合成数据作为元训练集进行元训练，将真实数据作为元测试集进行元测试，训练过程如图3所示。

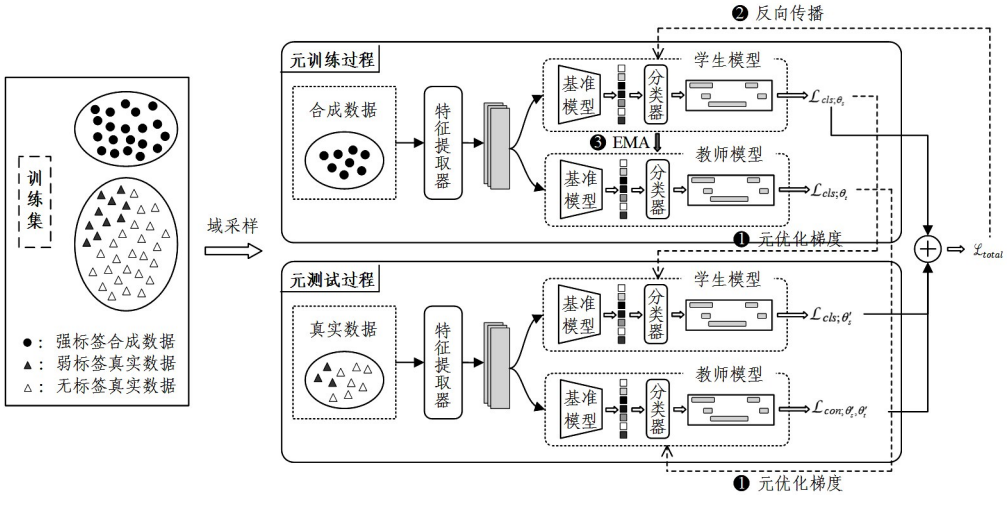


图3 基于元学习的半监督声音事件检测方法框架

Fig. 3 Framework of SSED based on meta learning

3.2.1 元训练

将元训练集的合成数据 x_s 分别送入学生模型 $f_{\theta_s}(x_s)$ 和教师模型 $f_{\theta_t}(x_s)$ 中，利用合成数据的强标签信息 y_s 进行监督训练，独立地更新学生和教师网络的参数 θ_s 和 θ_t。其计算式如式(5)所示：

$$\begin{aligned}\theta_s' &= \theta_s - lr \cdot \nabla_{\theta_s} \mathcal{L}_{\text{train\_s}} \\ \theta_t' &= \theta_t - lr \cdot \nabla_{\theta_t} \mathcal{L}_{\text{train\_t}}\end{aligned} \tag{5}$$

其中，$\mathcal{L}_{\text{train\_s}}$ 和 $\mathcal{L}_{\text{train\_t}}$ 分别是通过式(1)计算的学生和教师模型的分类损失，$\nabla_{\theta_s}\mathcal{L}_{\text{train\_s}}$ 和 $\nabla_{\theta_t}\mathcal{L}_{\text{train\_t}}$ 分别表示当前训练步中学生和教师模型参数关于分类损失的梯度。通过梯度下降，分别更新学生和教师网络模型的参数，得到 θ_s' 和 θ_t'。需要注意的是，在元训练集上得到的 θ_s' 和 θ_t' 仅用于让模型学习合成数据的知识，最终用于测试模型在元测试集上的性能。

3.2.2 元测试

在元训练完成后，学生模型和教师模型学习到了元训练集(合成数据)中的知识，接下来需要测试模型在元测试集(真实数据)上的性能。具体来说，将元测试集中的弱标签真实数据 x_w 送入学生模型 $f_{\theta_s'}(x_w)$ 中，利用少量真实数据的弱标签信息 y_w 通过式(1)计算学生模型分类损失 $\mathcal{L}_{\text{test\_s}}$；同时，为了有效利用元测试集中的无标签真实数据 x_u，此时通过式(2)仅在弱标签和无标签的真实数据上计算学生模型 $f_{\theta_s'}(x_u)$ 和教师模型 $f_{\theta_t'}(x_u)$ 之间的一致性损失 $\mathcal{L}_{\text{con}}$。综合考虑分类损失和一致性损失，更新学生模型参数 θ_s，改进式(3)为式(6)：

$$\mathcal{L}_{\text{total}} = \beta \mathcal{L}_{\text{train\_s}} + (1-\beta)\mathcal{L}_{\text{test\_s}} + \lambda(t)\mathcal{L}_{\text{con}} \tag{6}$$

其中，β 为训练权重超参。在这个步骤中，通过 $\mathcal{L}_{\text{total}}$ 实际更新学生模型参数 θ_s，如式(7)所示：

$$\theta_s \leftarrow \theta_s - lr \cdot \nabla_{\theta_s} \mathcal{L}_{\text{total}} \tag{7}$$

教师模型通过式(4)计算学生模型的指数移动平均并对其进行更新。

使用训练数据训练模型完成之后进入评估阶段，评估模型在测试数据上的性能。测试数据独立于训练数据，由强标签真实数据组成。

4 实验设置

4.1 数据集

本文采用半监督声音事件检测主流的 DCASE2021 任务 4 数据集[31]进行对比实验。数据集中共包含 10 种声音事件，分别是：说话声(Speech)、狗叫(Dog)、猫叫(Cat)、警报声(Alarm/bell/ringing)、碗碟声(Dishes)、煎炸声(Frying)、搅拌机声(Blender)、水流声(Running water)、吸尘器声(Vacuum cleaner)和电动剃须刀声(Electric shaver/toothbrush)。数据集包含训练集、验证集和测试集。其中，训练集由 10000 条强标签合成音频样本、1578 条弱标签真实音频样本和 14412 条无标签真实音频样本组成；验证集包含 1168 条强标签真实音频样本，用于模型调参；测试集包含 692 条强标签真实音频样本。所有音频样本的持续时间均不超过 10 s。其中，所有真实样本都是从 AudioSet[32]中采集获取，而合成样本则使用 SCAPER[33]工具包生成，以模拟真实样本的数据分布。

4.2 对比方法

目前，主流的半监督声音事件检测方法主要基于 MT 框架构建，这些方法使用的基准模型主要分为 CRNN(Convolutional Recurrent Neural Network)和 Conformer 两类。这些方法都需要获得帧级别(Frame-level)和句级别(Clip-level)的预测结果，分别用于一段音频的逐帧事件定位和整段事件的分类。

基于 CRNN 模型的方法通过聚合帧级别的预测结果得到句级别的预测结果，间接地对帧级别信息进行约束，通常具有更好的事件定位性能；而基于 Conformer 模型的方法通过额外增加一个输入向量的方式直接学习句级别的知识，即帧级别信息与句级别信息之间相对独立，通常具有更好的事件分类性能。

为了验证所提学习范式 MMT 的有效性和对于模型的可拓展性，本文将其与当前基于 CRNN 模型的先进方法[34-39]和基于 Conformer 模型的先进方法[25,40-41]进行对比，对比方法的总结如表 1 所列。其中 MMT-CRNN 和 MMT-Conformer 分别是基于 MMT 学习范式下使用 CRNN 和 Conformer 构建的方法。

表 1 对比方法的总结

Table 1 Summary of the compared methods

方法名称	基准模型	框架
Baseline21[34]	CRNN	MT
TBFL[35]	CRNN	MT
Coherence[36]	CRNN	MT
FDY[37-38]	CRNN	MT
BDS[39]	CRNN	MT
ConformerSED[40]	Conformer	MT
CNN-Trans[41]	Conformer	MT
Joint-Former[25]	Conformer	MT
MMT-CRNN	CRNN	MMT
MMT-Conformer	Conformer	MMT

4.3 评价指标

为了客观地比较实验结果，本文使用 DCASE 官方提供的两个指标作为衡量模型性能的参考依据。这两个指标分别是基于事件的 F1 分数指标[42](Event-based F1 Score，F1)和复合声音检测指标[43](Polyphonic Sound Detection Score，PSDS)。

基于事件的 F1 分数旨在比较系统输出和参考标签文件中每个事件实例起止时间的差异。具体计算式如式(8)所示：

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{8}$$

其中，P(Precision)表示精确率，即真阳性样本在所有被预测为阳性的样本中所占的比例；R(Recall)表示召回率，即真阳性样本在所有预测正确的样本中所占的比例。具体计算式如式(9)所示：

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{9}$$

其中，TP 表示真阳性样本数，即预测的类别正确且事件实例在参考值误差范围(Collar)内的样本数；FP 表示假阳性样本数，即在参考值范围内没有与预测类别相符的样本数；FN 表示假阴性样本数，即没有在参考值范围内检测到相应事件的样本数。

F1 分数强调事件实例发生的起止时间，而 PSDS 指标认为人工标注实例的起止时间戳具有一定的主观性，因此它从

另一种角度进行系统性能评估，使得系统对于具有主观性的标签更加鲁棒。具体地，其通过重新定义 TP 和 FP 的方式，允许间断发生的事件类以同一个实例的形式输出。例如，间断发生的多个狗叫实例也可以看作是一个输出实例。从实际用户体验出发，考虑同一音频内目标事件类别偏差对系统造成的影响。PSDS 指标的计算过程十分复杂，概括来说，首先 PSDS 通过 DTC 过滤器(Detection Tolerance Criterion)获得 FP，即筛选出系统输出实例的持续时间与这段持续时间内命中对应参考实例时间的比值需要大于阈值 ρ_{DTC} 的所有系统输出实例，不满足该条件的系统输出实例组成新的 FP 集合，计算这些 FP 集合的数量与所有系统输出实例时间的比值得到 FP 率(FP Rate，FPR)；其次 PSDS 通过 GTC 过滤器(Ground Truth Intersection Criterion)从 DTC 过滤器筛选出的这些系统输出实例中，筛选出参考实例的持续时间与这段时间内命中对应系统输出实例时间的比值需要大于阈值 ρ_{GTC} 的所有参考实例，组成新的 TP 集合，计算这些 TP 集合的数量与所有参考实例时间的比值得到 TP 率(TP Rate，TPR)，并通过计算在各类别上 TP 率的平均值和标准差得到有效 TP 率(Effective TP Rate，eTPR)；接着，定义 CTTC 过滤器(Cross-Trigger Tolerance Criterion)从新的 FP 集合中筛选出所有因目标事件实例混淆导致识别错误的所有系统输出实例，称为 CT 集合，通过计算 CT 集合的数量与被混淆的参考实例的时间的比值得到 CT 率(CT Rate，CTR)，并通过 FPR 和 CTR 的加权计算得到有效 FP 率(Effective FP Rate，eFPR)；最后 PSDS 定义为关于 eTPR 和 eFPR 的 ROC 曲线。本文遵循 DCASE 官方设置，定义 PSDS1 和 PSDS2 指标。前者侧重于事件定位性能，即 ρ_{DTC} 和 ρ_{GTC} 均设置为高阈值 0.7；后者侧重于事件分类性能，即 ρ_{DTC} 和 ρ_{GTC} 均设置为低阈值 0.1；两者 ρ_{CTTC} 均设置为 0.3。

4.4 实验环境及参数设置

实验使用 PyTorch 深度学习框架，显卡型号为 NVIDIA GeForce RTX 3080。在模型训练阶段，采用了 Rectified Adam 优化器[44]，学习率设置为 0.001，并使用了 StepLR 学习率调整器。每迭代 10000 步，学习率衰减至原来的 0.01，总的训练步数为 30000 步。

对于基于 CRNN 模型的方法和基于 Conformer 模型的方法，遵循 DCASE2021 官方基线[45]和 ConformerSED[40]的配置，并设置了不同的每批次训练的数据量。具体来说，基于 CRNN 的方法每批次训练的数据量为 48，其中包括 12 条合成数据、12 条弱标签真实数据和 24 条无标签真实数据；而基于 Conformer 模型的方法每批次训练的数据量为 128，其中包括 32 条合成数据，32 条弱标签真实数据和 64 条无标签真实数据。

在模型训练过程中，将 FBank 特征序列作为音频样本的输入。首先，将 44.1kHz 的音频样本重采样到 16kHz。为了保证对比实验的公平性，基于 CRNN 模型的方法和基于 Conformer 模型的方法采用了不同的特征提取参数设置。对于基于 Conformer 的模型，遵循 ConformerSED[40]的配置，使用窗口大小为 1024 的汉明窗对重采样后的样本进行分帧，帧移为 323，最后对每一帧样本进行快速傅里叶变换，通过 64 个通道的梅尔滤波器组再取对数，得到 496×64 维的 FBank 特征；而对于基于 CRNN 的模型，采用 DCASE2021 官方配置[45]，汉明窗大小为 2048，帧移为 256，梅尔滤波器组设为 128 个。

5 实验结果与分析

5.1 对比实验

5.1.1 与基线的对比

为了充分验证所提出的学习范式 MMT 的有效性，本小节使用 DCASE2021 数据集进行方案验证。首先，选择 DCASE2021 的官方基准方法 Baseline21[34]作为基线并与本文方法进行对比。Baseline21 基于 MT 框架，采用 CRNN 模型进行构建。本文提出的 MMT-CRNN 基于 MMT 范式，使用相同配置的 CRNN 模型进行模型训练。

实验结果如表 2 中的 Baseline21 和 MMT-CRNN 所示。所提方法 MMT-CRNN 降低了模型在验证集上过拟合的风险，并显著提升了模型在测试集上的音频事件分类与定位效果。具体来说，相较于基线 Baseline21，MMT-CRNN 在测试集上的 F1，PSDS1 和 PSDS2 指标上分别提升了 3.6%，2.2% 和 0.6%。

表 2 基于 MMT 的方法与基于 MT 的先进方法在 DCASE2021 数据集上的实验结果对比

Table 2 Comparison of experimental results between MMT-based method and MT-based advanced method on DCASE2021 dataset

(%)

对比方法	DCASE2021 验证集			DCASE2021 测试集		
	F1	*PSDS1*	*PSDS2*	*F1*	*PSDS1*	*PSDS2*
Baseline21[34]	38.5	24.8	43.1	40.5	33.4	53.6
TBFL[35]	—	—	—	41.3	—	—
Coherence[36]	—	—	—	43.5	—	—
FDY[37-38]	—	—	—	—	31.7	**54.3**
BDS[39]	**38.6**	**25.8**	**43.9**	43.9	34.5	54.2
MMT-CRNN	37.7	24.3	42.3	**44.1**	**35.6**	54.2
CNN-Trans[41]	—	—	—	—	29.2	55.0
ConformerSED[40]	41.4	25.8	45.0	42.4	29.7	52.0
Joint-Former[25]	41.9	**26.8**	45.1	44.2	**33.9**	55.1
MMT-Conformer	**42.0**	25.7	**47.9**	**46.1**	32.9	**61.8**

注："—"表示对应指标结果未公布，最优结果用加粗表示。

5.1.2 与当前先进方法的对比

此外，进一步将本文提出的方法与当前半监督声音事件检测领域中的其他先进方法进行对比，这些方法均基于 MT 框架进行构建。根据使用模型的不同，将它们分为两类：一类是基于 CRNN 模型的方法，包括 TBFL[35]，Coherence[36]，FDY[37-38]和 BDS[39]；另一类是基于 Conformer 模型的方法，包括 ConformerSED[40]，CNN-Trans[41] 和 Joint-Former[25]。本节对最新方法[25,39-40]中的模型进行复现，并给出验证集和测试集上所有指标的结果。

实验结果如表 2 所列。本文方法相较于上述半监督声音事件检测先进方法表现出显著的性能优势。具体而言，在基于 CRNN 模型的方法中，与 TBFL，Coherence 和 BDS 方法相比，MMT-CRNN 在 F1 指标上分别提升了 2.8%，0.8% 和 0.2%；与 FDY 和 BDS 的方法相比，MMT-CRNN 在 PSDS1 指标上分别提升了 3.9% 和 1.1%，在 PSDS2 指标上几乎与

FDY 和 BDS 的性能持平。在基于 Conformer 模型的方法中，MMT-Conformer 在几乎不损失定位性能的情况下，分类性能相较于先进方法取得了显著的性能提升。具体而言，MMT-Conformer 相较于 CNN-Trans[41]，在 PSDS1 和 PSDS2 指标上分别提升了 3.7%和 6.8%；相较于 ConformerSED，在 F1，PSDS1 和 PSDS2 指标上分别提升了 3.7%，3.2%和 9.8%；相较于 Joint-Former，在 F1 和 PSDS2 指标上分别提升了 1.9%和 6.7%。

实验结果表明，所提出的框架 MMT 在考虑数据域差异、充分利用大量合成数据以提升模型泛化性能方面表现出了可行性，并且对于模型具有一定的可拓展性。在 DCASE2021 数据集的验证实验中，MMT-CRNN 相较于官方基线 Baseline21 表现出更好的泛化性能，尤其在测试集上的 F1，PSDS1 和 PSDS2 指标上均取得了显著提升。

相较于其他先进方法，MMT 在性能上取得了显著的优势，具体体现为在 F1，PSDS1 和 PSDS2 指标上的提升。

综合实验结果可以得出：MMT 在半监督声音事件检测任务中具有潜力，能够有效应对域差异，提升模型泛化性能，并在大量合成数据的支持下取得显著的性能优势。

5.2 消融实验

为了验证提出的基于元学习的学习范式对于合成域泛化到真实域进而提升模型检测性能的有效性，本节在 DCASE2021 数据集上设计了两组消融实验，从两个角度探究所提学习范式 MMT 对模型检测性能的影响。

1）探究在 MMT 学习范式下合成数据量与模型检测性能的相关性。本文提出的学习范式 MMT 旨在使模型显式地暴露在域差异环境下，从而有效利用大量易获取的合成数据，充分学习与任务相关的知识来进一步提升声音事件检测性能。具体而言，合成数据量越大，模型习得的知识越丰富，越有利于事件检测。为了验证在 MMT 学习范式下合成数据量与检测性能的正相关性，本节在 DCASE2021 训练集上基于所提方法 MMT-CRNN 进行模型训练，对于训练集中的 10000 条合成数据，分别设置为原数据量的 1/5(2000 条)、2/5(4000 条)、3/5(6000 条)、4/5(8000 条)和 1(10000 条)进行训练。最后，在 DCASE2021 测试集上对 MMT-CRNN 进行性能评估。

实验结果如表 3 所列。随着训练集中合成数据量的增多，MMT-CRNN 有相对稳定的性能提升。分析实验结果可知，数据量的较大差异导致模型在进行元训练的过程中无法充分学习到合成数据中事件检测的相关知识。换言之，源域数据的明显不足可能导致模型无法通过元学习捕捉到重要的通用特征，从而导致性能下降。

表 3 在 MMT 学习范式下合成数据量与模型检测性能的相关性

Table 3 Correlation between synthetic data volume and model detection performance in MMT learning paradigm

合成数据量/条	DCASE2021 测试集/%		
	*F*1	*PSDS*1	*PSDS*2
2000	38.1	24.8	35.3
4000	39.4	32.1	49.5
6000	38.4	31.5	48.3
8000	41.2	33.5	51.3
10000	**44.1**	**35.6**	**54.2**

2）探究所提范式 MMT 中元梯度和一致性正则技术对模型检测性能的影响。模型通过元梯度进行两次梯度更新调整来学习合成数据到真实数据的域泛化，通过一致性正则化技术对真实域中无标签数据到弱标签数据进行进一步的标签泛化，从而显著提升了模型的事件检测性能。因此，本小节以声音事件检测任务中主流的模型 CRNN 和 Conformer 分别作为基线设计消融实验，如表 4 所列。具体来说，首先不使用 MMT 而直接使用 CRNN 和 Conformer 在有标签数据上进行模型训练；然后，基于元梯度构建 Meta-CRNN 和 Meta-Conformer 模型，将模型明确地暴露在有标签数据的域差异环境下；最后，使用 MMT 范式构建所提方法 MMT-CRNN 和 MMT-Conformer 模型。相较于 Meta-CRNN 和 Meta-Conformer，基于 MMT 的模型进一步通过一致性正则充分利用了无标签真实数据。模型均在 DCASE2021 测试集上进行性能评估。

表 4 消融实验设计

Table 4 Design of ablation experiment

模型	元梯度	一致性正则
CRNN		
Meta-CRNN	√	
MMT-CRNN	√	√
Conformer		
Meta-Conformer	√	
MMT-Conformer	√	√

消融实验结果如表 5 所列。可以看出，无论是 CRNN 还是 Conformer 模型，加上本文所提出的基于元学习的声音事件检测学习范式后都获得了显著的性能提升。具体来说，Meta-CRNN 相较于 CRNN 在 F1，PSDS1 和 PSDS2 指标上分别提升了 2.0%，3.6%和 2.2%，MMT-CRNN 相较于 CRNN 在 F1，PSDS1 和 PSDS2 指标上分别提升了 4.9%，6.0%和 3.3%；Meta-Conformer 相较于 Conformer 在 F1，PSDS1 和 PSDS2 指标上分别提升了 1.3%，2.3%和 1.1%，MMT-Conformer 相较于 Conformer 在 F1，PSDS1 和 PSDS2 指标上分别提升了 8.8%，6.4%和 12.1%。此外，在使用元梯度进行域泛化的基础上，通过一致性正则技术充分利用无标签数据可以进一步提升模型的检测性能。具体来说，MMT-CRNN 相较于 Meta-CRNN 在 F1，PSDS1 和 PSDS2 指标上分别提升了 2.9%，2.4%和 1.1%；MMT-Conformer 相较于 Meta-Conformer 在 F1，PSDS1 和 PSDS2 指标上分别提升了 7.5%，4.1%和 11.0%。

表 5 所提范式 MMT 中元梯度和一致性正则技术对模型检测性能的影响

Table 5 Impact of meta-gradient and consistency regularization techniques on model performance in MMT paradigm

(%)

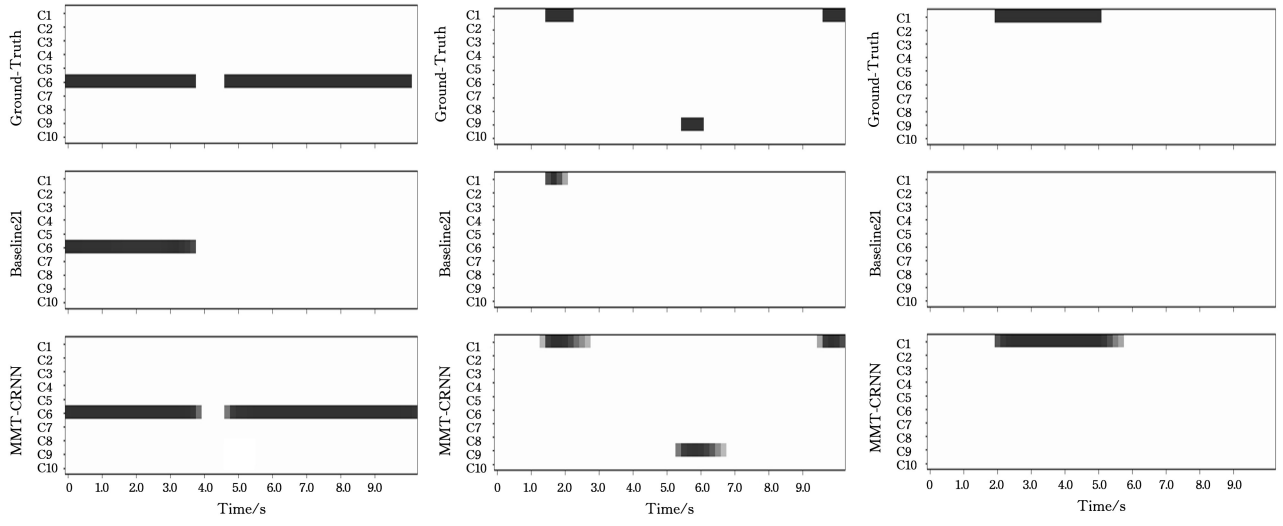
模型	DCASE2021 测试集		
	*F*1	*PSDS*1	*PSDS*2
CRNN	39.2	29.6	50.9
Meta-CRNN	41.2	33.2	53.1
MMT-CRNN	**44.1**	**35.6**	**54.2**
Conformer	37.3	26.5	49.7
Meta-Conformer	38.6	28.8	50.8
MMT-Conformer	**46.1**	**32.9**	**61.8**

上述实验结果充分证明了本文提出的基于元学习的半监督声音事件检测学习范式 MMT 的有效性。

5.3 可视化结果分析

本节给出所提方法 MMT-CRNN 与 Baseline21(即基于 MT 的 CRNN)的可视化结果对比,如图 4 所示。图 4 中的每张子图从上至下分别是真实标签(Ground Truth)、Baseline21 和 MMT-CRNN 随时间变化在对应事件类别上的定位结果。纵轴表示数据集中的 10 个声音事件类,C1—C10 分别代表事件类别警报声、搅拌机声、猫叫、碗碟声、狗叫、电动剃须刀声、煎炸声、水流声、说话声和吸尘器声;横轴表示音频事件序列,以秒为单位。

如图 4 所示,本文从 DCASE2021 测试集中选取了 3 条真实数据样本进行可视化结果分析,样本编号分别为 e8jhGkgHG44_0_10,6CLqqCp8Jfo_138_148 和 hFhsNDQ1mEo_0_5,对应可视化结果分别为图 4(a)—图 4(c)。在图 4(a)中,所提方法 MMT-CRNN 成功定位出 Baseline21 漏检的第 4.5 秒至第 10 秒的长时电动剃须刀声;在图 4(b)中,MMT-CRNN 成功定位出 Baseline21 漏检的短时警报声和说话声;在图4(c)中,Baseline21 误将警报声识别为猫叫,而 MMT-CRNN 在这种情况下也可以成功定位出警报声在该音频中的起止时间。通过对比实验中的可视化结果可知,MMT 框架在声音事件检测中表现出更好的分类和定位性能。



(a)[File ID]e8jhGkgHG44_0_10.wav (b)[File ID]6CLqqCp8Jfo_138_148.wav (c)[File ID]hFhsNDQ1mEo_0_5.wav

图 4 DCASE2021 测试集中的测试样本的事件检测可视化结果

Fig. 4 Event detection visualization results of test samples in DCASE2021 test set

6 讨论

先前的半监督声音事件检测方法大多未考虑合成数据与真实数据域之间天然存在的分布差异,这种分布差异会限制模型从合成域泛化到真实域的能力,进而影响模型在真实数据上的检测性能。相反,本文方法 MMT 将模型显式地暴露在域差异环境下,使其能够更好地适应不同领域的数据从而提高泛化能力,这为提升半监督声音事件检测性能提供了新的思路和方法,可以更有效地利用大量的标签易获取的合成数据监督模型,从合成数据中学习有效知识,在提升检测性能的同时降低模型过拟合的可能性。

尽管相较于先进方法,MMT 表现出了显著的性能优势,但也存在不足之处,主要在于 MMT 仍依赖不同数据域的有标签数据进行训练,没有考虑对无标签真实数据进行更充分的分析与利用;此外,模型需要计算二阶导元梯度,相比于平均教师框架增加了模型计算的复杂度。

未来改进方向可以通过在数据域上先进行无监督聚类迭代获取标签的方式更新数据域标签,避免产生人工标注数据域的成本。此外,可以考虑利用 MMT 学习真实域上有标签和无标签数据之间存在的分布差异,进一步校准模型优化方向。

结束语 针对半监督声音事件检测任务中合成数据和真实数据之间存在的域差异影响模型泛化性能的问题,本文提出了一种基于元学习的半监督声音事件检测学习范式 MMT,利用元学习指导模型从合成域到真实域的梯度更新,使模型感知并学习域泛化知识,提高模型泛化能力,从而提升检测性能。在 DCASE2021 数据集上的实验结果充分证明了所提方法的有效性和泛化性。MMT 尽管存在一定的不足,但仍表现出不错的检测性能潜力。作为一种可用于跨域声音事件检测的通用架构,未来的研究工作考虑将该方法应用到具有更多种类的声音事件和环境条件的复杂场景。

参 考 文 献

[1] ROVETTA S, MNASRI Z, MASULLI F, et al. Anomaly Detection Based on Interval-Valued Fuzzy Sets: Application to Rare Sound Event Detection[C]// Proceedings of the 13th International Workshop on Fuzzy Logic and Applications. Vietri sul Mare: Springer Press, 2021: 1-8.

[2] NERI M, BATTISTI F, NERI A, et al. Sound Event Detection for Human Safety and Security in Noisy Environments[J]. IEEE Access, 2022, 10: 134230-134240.

[3] PANDEYA Y R, BHATTARAI B, LEE J. Visual Object Detector for Cow Sound Event Detection[J]. IEEE Access, 2020, 8: 162625-162633.

[4] GAO L J, MAO Q R. Environment-assisted Multi-task Learning

for Polyphonic Acoustic Event Detection[J]. Computer Science, 2020,47(1):159-164.

[5] SERIZEL R,TURPAULT N,EGHBAL Z H,et al. Large-Scale Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments[C]// Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events. Surrey:IEEE Press,2018:1-5.

[6] TARVAINEN A,VALPOLA H. Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach:MIT Press,2017:1195-1204.

[7] TURPAULT N,SERIZEL R,SHAH A P,et al. Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis[C]// Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events. New York:IEEE Press,2019:1-6.

[8] HU Y,ZHU X J,LI Y L,et al. A Multi-Grained Based Attention Network for Semi-Supervised Sound Event Detection[C]// Proceedings of the International Conference on INTERSPEECH. Incheon:ISCA,2022:1531-1535.

[9] SHAO N,LOWEIMI E,LI X F. RCT:Random Consistency Training for Semi-Supervised Sound Event Detection[C]// Proceedings of the International Conference on INTERSPEECH. Incheon:ISCA,2022:1541-1545.

[10] NAM H,KIM S H,KO B Y,et al. Frequency Dynamic Convolution:Frequency-Adaptive Pattern Recognition for Sound Event Detection[C]// Proceedings of the International Conference on Iterspeech. Incheon:ISCA,2022:2763-2767.

[11] PARK S,KOTHINTI S R,ELHILALI M. Temporal Coding with Magnitude-Phase Regularization for Sound Event Detection [C]// Proceedings of the International Conference on INTERSPEECH. Incheon:ISCA,2022:1536-1540.

[12] YANG L P,HAO J Y,GU X H,et al. Sound Event Detection with Audio Tagging Consistency Constraint CRNN[J]. Journal of Electronics & Information Technology,2022,44(3):1102-1110.

[13] YANG S Z,ZHANG L,WANG J H,et al. Review of Sound Event Detection[J]. Journal of Guangxi Normal University (Natural Science Edition),2023,41(2):1-18.

[14] YANG L P,HAO J Y,HOU Z W,et al. Two-Stage Domain Adaptation for Sound Event Detection[C]// Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events. Tokyo:IEEE Press,2020:230-234.

[15] ZHENG X,SONG Y,DAI L R,et al. An Effective Mutual Mean Teaching Based Domain Adaptation Method for Sound Event Detection[C]// Proceedings of the International Conference on INTERSPEECH. Brno:ISCA,2021:556-560.

[16] HUISMAN M,VAN RIJN J N,PLAAT A. A Survey of Deep Meta-Learning[J]. Artificial Intelligence Review,2021,54(6):4483-4541.

[17] WEI Q Y,YU L Q,LI X H,et al. Consistency-Guided Meta-Learning for Bootstrapping Semi-Supervised Medical Image Segmentation[C]// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Vancouver:Springer Press,2023:183-193.

[18] XU H,XIE H T,TAN Q F,et al. Meta Semi-Supervised Medical Image Segmentation with Label Hierarchy[J]. Health Information Science and Systems,2023,11(1):26.

[19] LI J N,WONG Y K,ZHAO Q,et al. Learning to Learn from Noisy Labeled Data[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach:IEEE Press,2019:5051-5059.

[20] ALGAN G,ULUSOY I. MetaLabelNet:Learning to Generate Soft-Labels from Noisy-Labels[J]. IEEE Transactions on Image Processing,2022,31:4352-4362.

[21] LI J Z,SUN H L. Correct Twice at Once:Learning to Correct Noisy Labels for Robust Deep Learning[C]// Proceedings of the 30th ACM International Conference on Multimedia. Lisbon:ACM Press,2022:5142-5151.

[22] ZHU W T,LIU W,LIANG S S,et al. Variational Continuous Bayesian Meta-Learning Based Algorithm for Recommendation [J]. Computer Science,2023,50(7):66-71.

[23] BAI J,GENG X Y,YI L,et al. Improved Feature Interaction Algorithm Based on Meta-learning[J/OL]. https://www.jsjkx.com/CN/article/openArticlePDF.jsp? id=22016.

[24] WEI W,ZHU H,BENETOS E,et al. A-CRNN:A Domain Adaptation Model for Sound Event Detection[C]// Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing. Barcelona:IEEE Press,2020:276-280.

[25] GAO L J,MAO Q R,DONG M. Joint-Former:Jointly Regularized and Locally Down-sampled Conformer for Semi-Supervised Sound Event Detection[C]// Proceedings of the International Conference on INTERSPEECH. Dublin: ISCA, 2023: 2753-2757.

[26] HOSPEDALES T,ANTONIOU A,MICAELLI P,et al. Meta-Learning in Neural Networks:A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5149-5169.

[27] YAO H X,WU X,TAO Z Q,et al. Automated Relational Meta-Learning[C]// Proceedings of the 8th International Conference on Learning Representations. Virtual:Ithaca,2020:1-19.

[28] FINN C,ABBEEL P,LEVINE S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks[C]// Proceedings of the International Conference on Machine Learning. Sydney: ACM Press,2017:1126-1135.

[29] ZHOU K,LIU Z,QIAO Y,et al. Domain Generalization:A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,45(4):4396-4415.

[30] ZENG C,WANG X,MIAO X X,et al. Improving Generalization Ability of Countermeasures for New Mismatch Scenario by Combining Multiple Advanced Regularization Terms[C]// Proceedings of the International Conference on INTERSPEECH. Dublin:ISCA,2023:1998-2002.

[31] SERIZEL R,TURPAULT N,SHAH A,et al. Sound Event Detection in Synthetic Domestic Environments[C]// Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing. Barcelona:IEEE Press,2020:86-90.

[32] GEMMEKE J F, ELLIS D P W, FREEDMAN D, et al. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans: IEEE Press, 2017: 776-780.

[33] SALAMON J, MACCONNELL D, CARTWRIGHT M, et al. Scaper: A Library for Soundscape Synthesis and Augmentation [C]// Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz: IEEE Press, 2017: 344-348.

[34] RONCHINI F, SERIZEL R. A Benchmark of State-of-the-Art Sound Event Detection Systems Evaluated on Synthetic Soundscapes [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE Press, 2022: 1031-1035.

[35] PARK S, ELHILALI M. Time-Balanced Focal Loss for Audio Event Detection [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE Press, 2022: 311-315.

[36] KOTHINTI S, ELHILALI M. Temporal Contrastive-Loss for Audio Event Detection [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE Press, 2022: 326-330.

[37] NAM H, KIM S H, KO B Y, et al. Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection [C]// Proceedings of the International Conference on INTERSPEECH. Incheon: ISCA, 2022: 2763-2767.

[38] WANG J, YAO P, DENG F, et al. NAS-DYMC: NAS-Based Dynamic Multi-Scale Convolutional Neural Network for Sound Event Detection [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE Press, 2023: 1-5.

[39] LIN W C, BONDI L, GHAFFARZADEGAN S. Background Domain Switch: A Novel Data Augmentation Technique for Robust Sound Event Detection [C]// Proceedings of the International Conference on INTERSPEECH. Dublin: ISCA, 2023: 326-330.

[40] MIYAZAKI K, KOMATSU T, HAYASHI T, et al. Conformer-Based Sound Event Detection with Semi-Supervised Learning and Data Augmentation [C]// Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events. Tokyo: IEEE Press, 2020: 101-104.

[41] WAKAYAMA K, SAITO S. CNN-Transformer with Self-Attention Network for Sound Event Detection [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE Press, 2022: 806-810.

[42] MESAROS A, HEITTOLA T, VIRTANEN T. Metrics for Polyphonic Sound Event Detection [J]. Applied Sciences, 2016, 6(6): 162.

[43] BILEN Ç, FERRONI G, TUVERI F, et al. A Framework for the Robust Evaluation of Sound Event Detection [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE Press, 2020: 61-65.

[44] LIU L Y, JIANG H M, HE P C, et al. On The Variance of the Adaptive Learning Rate and Beyond [C]// Proceedings of International Conference on Learning Representations. Virtual: Ithaca, 2020: 1-14.

[45] TURPAULT N, SERIZEL R. Training Sound Event Detection on A Heterogeneous Dataset [C]// Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events. Tokyo: IEEE Press, 2020: 1-5.

SHEN Yaxin, born in 1999, postgraduate. Her main research interests include multimedia intelligent analysis and so on.

MAO Qirong, born in 1975, professor, Ph.D supervisor, is a member of CCF (No. 17370S). Her main research interests include multimedia intelligent analysis and emotional computing.

(责任编辑:何杨)