软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007154]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于半监督学习的长尾时序动作检测*

王雨虹, 武港山, 王利民

(计算机软件新技术国家重点实验室 (南京大学), 江苏 南京 210023)

通信作者: 王利民, E-mail: lmwang@nju.edu.cn

摘　要: 现实世界中的数据标签分布往往呈现长尾效应, 即少部分类别占据绝大多数样本, 时序动作检测问题也不例外. 现有的时序动作检测方法往往缺乏对少样本类别的关注, 即充分建模样本数量多的头部类别, 而忽视了样本数量少的尾部类别. 对长尾时序动作检测问题进行了系统的定义, 并针对长尾时序动作检测问题, 提出一种基于半监督学习的加权类别重平衡自训练方法, 充分利用现实世界中存在的大规模无标签数据, 来重平衡训练样本中的标签分布, 改善模型对尾部类别的拟合效果. 还针对时序动作检测任务, 提出一种伪标签损失加权方法, 使模型训练更加稳定. 在 THUMOS14 和 HACS Segments 数据集上进行实验, 并分别利用 THUMOS15 数据集和 ActivityNet1.3 数据集中的视频样本来构成相应的无标签数据集. 此外, 还针对视频审核应用需求, 收集 Dance 数据集, 包括 35 个动作类别、6 632 个有标签视频和 13 264 个无标签视频, 并保留数据分布显著的长尾效应. 使用多种基线模型, 在 THUMOS14、HACS Segments 和 Dance 数据集上进行实验. 实验结果表明, 所提出的加权类别重平衡自训练方法可以提高模型对尾部动作类别的检测效果, 并且能应用于不同的基线时序动作检测模型提升其性能.

关键词: 视频分析; 时序动作检测; 深度长尾学习; 半监督学习

中图法分类号: TP18

中文引用格式: 王雨虹, 武港山, 王利民. 基于半监督学习的长尾时序动作检测. 软件学报. http://www.jos.org.cn/1000-9825/7154.htm

英文引用格式: Wang YH, Wu GS, Wang LM. Long-tailed Temporal Action Detection Based on Semi-supervised Learning. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7154.htm

Long-tailed Temporal Action Detection Based on Semi-supervised Learning

WANG Yu-Hong, WU Gang-Shan, WANG Li-Min

(State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: The label distribution in the real world often shows the long-tail effect, where a small number of categories account for the vast majority of samples. The temporal action detection problem is no exception. The existing temporal action detection methods often focus on the head categories with a large number of samples, while neglecting the few-sample categories. This study systematically defines the long-tail temporal action detection problem and proposes a weighted class-rebalancing self-training method (WCReST) based on a semi-supervised learning framework. WCReST makes full use of the large-scale unlabeled data that exists in the real world to rebalance the label distribution in the training samples to improve the model's fit for the tail categories. Additionally, a pseudo-label loss weighting method is proposed for the temporal action detection task to enhance the stability of model training. Experiments are conducted on the THUMOS14 and HACS Segments datasets, using video samples from the THUMOS15 and ActivityNet1.3 datasets to form corresponding unlabeled datasets. In addition, the Dance dataset is collected to meet the application requirements of video review, which includes 35 action categories, 6 632 labeled videos, and 13 264 unlabeled videos, preserving the significant long-tail effect in data distribution. A variety of baseline models are used to conduct experiments on the THUMOS14, HACS Segments, and Dance datasets. The results demonstrate that the proposed WCReST can improve the model's detection performance on tail action categories and can be applied to

* 基金项目: 科技创新 2030—"新一代人工智能"重大项目 (2022ZD0160900); 国家自然科学基金 (62076119, 61921006)
收稿时间: 2023-08-11; 修改时间: 2023-09-07; 采用时间: 2023-12-22; jos 在线出版时间: 2024-07-17

different baseline temporal action detection models to enhance their performance.

Key words: video analysis; temporal action detection; deep long-tailed learning; semi-supervised learning

1 简 介

随着互联网技术和多媒体采集设备的迅速发展, 人们逐渐倾向于使用可承载信息容量更大的视频来记录生活, 视频数据量呈爆炸性增长. 更多视频数据的产生自然也带来了大量的应用场景, 视频理解技术在视频检索、安防监控、人机交互等领域都有着广阔的应用前景. 与图像相比, 视频数据增添了时序维度, 更为复杂多样, 对计算机算力需求更高. 此外, 未经剪辑的视频数据往往较长, 信息密度较低, 这就要求模型能够高效理解视频内容, 精确定位有价值的信息.

时序动作检测任务[1]是计算机视觉领域的经典问题, 如图 1 所示, 其研究目标是检测出一段未经剪辑的长视频中动作发生的区间 (包括开始时间和结束时间) 和动作类别. 与其他视觉相关任务一样, 时序动作检测任务的数据也往往呈现长尾分布, 即少部分头部动作类别具有较多的样本数量, 而大部分尾部动作类别则样本数量较少. 现有时序动作检测工作通常缺乏对长尾问题的关注, 动作检测模型倾向于拟合头部类别, 而缺少对尾部类别的建模. 然而在实际应用的过程中, 尾部类别可能具有与头部类别同等甚至更高的重要性. 比如在舞蹈视频审核时, 违规舞蹈动作出现的频率远低于正常舞蹈动作, 属于数据中尾部类别, 然而模型需要检测的正是这些尾部动作类别. 因此, 如何在具有长尾标签分布的数据上, 提高时序动作检测模型建模尾部类别的能力, 也有着极高的研究意义.

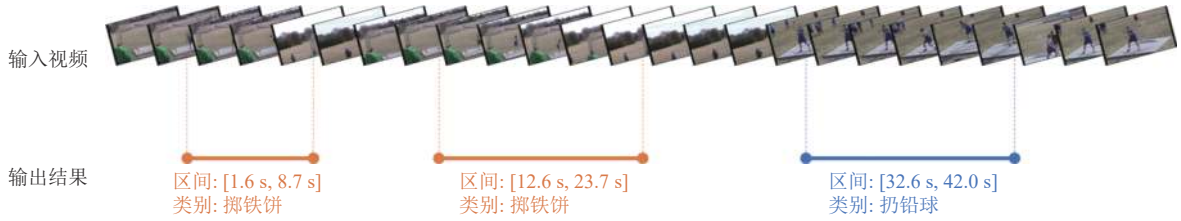


图 1 时序动作检测任务示意图

针对数据标签分布的长尾问题, 常见的处理方法有重采样、重加权、度量学习、解耦训练等. 本文则尝试采用半监督学习的方法来改善模型在长尾时序动作检测任务上的性能表现, 原因主要有以下两点: (1) 现实世界中存在海量的无标签数据, 收集这些数据的成本极为低廉, 然而对这些数据进行标注却需要耗费大量的人力和时间. 半监督学习方法可以充分利用这些无标签数据, 避免资源浪费. (2) 半监督学习方法较为通用, 可以嵌入到任何模型的训练过程中, 同时也可以和其他长尾学习方法结合, 进一步提升模型的性能表现.

本文针对时序动作检测任务的特点, 提出了一种基于半监督学习的加权类别重平衡自训练方法, 简称为 WCReST (weighted class-rebalancing self-training). 具体来说, 在每次迭代过程中, 首先会基于原有模型在无标签数据上生成伪标签, 然后根据类别重平衡采样策略来采样合适的视频样本加入原有的训练集中, 再在新数据集上训练得到新的模型. 其中, 类别重平衡采样策略的本质是, 根据训练集中的数据分布来采样新样本, 频繁采样出现次数少的尾部动作类别, 而较少采样出现次数多的头部动作类别. 通过这种方式, 来重新平衡训练集中的样本分布, 并保证伪标签的准确性. 特别地, 考虑到检测任务与分类任务之间的差异, 本文分别在片段级别和视频级别上为伪标签添加损失权重, 用以表示其在模型迭代训练过程中的可靠性, 从而增加迭代训练过程的稳定性.

本文的贡献主要有以下 3 点: (1) 本文首次关注时序动作检测任务中的长尾问题, 并使用半监督学习方法来提高模型对尾部类别的拟合能力, 设计了一种加权类别重平衡自训练 (WCReST) 方法, 在不断迭代训练模型的过程中, 重平衡数据集中的类别样本分布. 同时, 本文还针对时序动作检测任务特点, 创新性地提出了片段级别和视频级别的伪标签加权方法, 从而增加模型训练的稳定性. (2) 本文在 THUMOS14[1]和 HACS Segments[2]这两个时序动作检测数据集上进行了实验, 其对应的无标签数据集分别由 THUMOS15 数据集[3]和 ActivityNet1.3 数据集[4]构成.

实验结果证明了 WCReST 方法的有效性. (3) 本文面向舞蹈视频审核应用, 收集并整理了 Dance 数据集, 并保留了动作类别分布的显著长尾效应. 本文在 Dance 数据集上进行了详尽的对比实验, 证明 WCReST 方法可以提高时序动作检测模型在长尾数据上的性能, 具有很高的现实应用价值.

本文第 2 节简要介绍相关工作. 第 3 节详细描述本文提出的 WCReST 方法. 第 4 节介绍本文实验所使用的数据集, 并介绍 Dance 数据集的构建方法. 第 5 节展示并分析 WCReST 方法在数据集上的对比实验结果. 第 6 节进行总结与展望.

2 相关工作

随着多媒体技术的不断发展, 对长视频内容理解的需求逐渐增加, 时序动作检测任务逐渐成为计算机视觉领域的热点研究问题. 其研究目标是, 检测出未经剪辑的长视频中动作发生的开始时间、结束时间和动作类别, 在视频审核、智能监控、 虚拟现实等领域都有着广泛的应用.

较早的时序动作检测模型多采用两阶段框架[5−14], 将时序动作检测分为时序动作定位 (又名时序动作提名生成) 和动作提名分类这两个子任务. 具体来说, 时序动作定位任务负责定位动作的开始和结束时间点, 生成动作提名候选框; 动作提名分类任务则负责对上一阶段生成的动作提名片段进行分类, 预测其动作类别. 这些两阶段框架通过解构和拆分复杂的时序动作检测问题, 降低任务求解难度, 也使得算法模型更加灵活, 训练成本更低. 然而, 两阶段时序动作检测框架的计算效率往往较为低下, 且训练和推理过程复杂, 不能满足很多端到端的实际应用需求. 因此, 随着对时序动作检测领域研究的深入, 研究者们也提出了一系列单阶段时序动作检测方法.

单阶段时序动作检测方法[15−22]将定位动作边界任务和预测动作类别任务当作一个整体, 使用端到端模型对输入的未剪辑长视频中的动作进行定位和分类. SSAD[15] 受目标检测领域的 SSD 模型和 YOLO 模型的启发, 将基于时序卷积的单阶段结构引入了时序动作检测领域, 并采用多种时序卷积模型来提取多尺度的视频特征. R-C3D[16] 则参考目标检测模型 Faster R-CNN 的思路, 先生成时序提名, 再进行 3D 感兴趣区域池化, 最后进行动作分类和边界回归. TAL-Net[17] 同样受 Faster R-CNN 启发, 使用多尺度架构的感受野对齐来适应动作长度的巨大差异, 并显式扩展提名生成和动作分类的感受野来更好地利用时序上下文. AFSD[19] 则构建了第 1 个 anchor-free 的端到端时序动作检测框架, 提出边界池化操作和边界一致性对比学习方法, 来提取显著边界特征并保证其有效性. 随着 Transformer 在计算机视觉领域的逐渐普及, 不少工作也尝试将 Transformer 框架引入端到端时序动作检测领域. TadTR[20] 基于 Transformer 框架, 将所有动作实例定义为一组并行预测的动作标签-时序位置对, 自适应地提取时间上下文信息来进行时序动作检测. 参考 DETR[23] 框架, ReAct[21] 也使用带有动作查询的编码器-解码器结构, 并提出一种关系注意机制来指导解码器中查询间的自注意力计算, 还使用对比学习方法来提升分类头的准确性. ActionFormer[22] 使用多尺度 Transformer 编码器来关注不同时间长度的动作实例, 并使用轻量级卷积网络解码器来进行动作分类和边界回归. 单阶段的时序动作检测模型更为简洁和高效, 工程复杂度更低, 端到端的检测方式也使得模型更容易被复现和落地, 在现实生活中应用范围更加广阔.

和许多其他的视觉问题类似, 在实际应用中, 时序动作检测任务的数据大多也呈现长尾分布. 尽管目前关注长尾时序动作检测的研究较少, 但在其他视觉任务上, 研究者们已经对深度长尾问题进行了较为充分的研究, 并提出了一系列数据集和算法模型. 现有的长尾学习工作可以分为重采样、重加权、迁移学习、度量学习、解耦训练、半监督学习等多个类别. 重采样方法分为过采样[24] 和欠采样[25] 两种, 通过重复采样尾部类别或丢弃部分头部类别, 来重新平衡数据的类别分布. 重加权方法[26−30] 则通过自适应的方式, 对不同类别赋予不同的损失函数权重, 增加模型对尾部类别的关注. 而面向长尾问题的迁移学习方法[31,32] 则对头部类别和尾部类别分别训练不同的模型, 并且将在头部类别样本中学习到的知识迁移至尾部类别来帮助建模. 度量学习方法[33,34] 的研究思路是, 希望模型能够学习到更合适的特征嵌入, 从而能够更好地建模尾部类别样本的边缘. 解耦训练方法[35−38] 则将特征学习和分类器学习解耦, 使用两阶段方法来训练模型. 半监督长尾学习方法[39−42] 则利用现实存在的大量无标签数据来辅助训练模型, 采用自训练方法, 对无标签数据生成伪标签进而一起训练.

与其他方法相比, 基于半监督学习的长尾方法具有较强的通用性和可嵌入性, 且能够充分挖掘不平衡数据标签的有价值信息, 并帮助模型克服固有的训练标签偏差, 获得更好的性能. 在长尾图像分类问题上, 文献 [39] 首次尝试从半监督学习和自监督学习的角度去理解不均衡数据标签的价值. 文献 [41] 则提出了一个抑制一致性损失函数来抑制少数类的损失. 文献 [42] 则通过凸优化方法来改进原始生成的伪标签. 而 CReST[40] 则制定了一种基于类别重平衡原则的伪标签数据采样策略, 通过更频繁地采样尾部类别的伪标签样本数据, 来重新平衡数据的类别分布.

受 CReST[40] 启发, 本文也尝试将这种半监督自训练方法应用到长尾时序动作检测问题上来. 然而, 直接将其应用在时序动作检测任务上时, 会面临两个问题: (1) 与图片数据相比, 视频数据增添了时序维度, 计算机算力需求更高. 此外, 未经剪辑长视频的信息密度较低, 因此, 直接生成伪标签并用于网络自训练的方法效率较低, 且容易受到无效信息的干扰. (2) 与分类任务不同, 检测任务不仅需要识别动作的类别, 还要定位动作发生的区间. 因此, 一方面, 模型需要额外衡量伪标签在定位任务上的可靠性, 避免降低模型定位的精确性. 另一方面, 同一个输入视频可能生成多个动作片段伪标签, 如何在避免噪声干扰的同时, 利用这些伪标签片段中的有价值部分, 也是模型面临的一大难点.

针对以上这两点, 本文提出了一种面向长尾时序动作检测问题的加权类别重平衡自训练方法 WCReST. 不仅通过类别重平衡采样的方式来均衡数据集中的类别分布, 还分别在片段级别和视频级别上为伪标签添加可靠性权重, 提高训练效率, 也增加迭代训练的稳定性.

3 基于半监督学习的长尾时序动作检测方法

时序动作检测任务的研究目标是, 输入一段未经剪辑的长视频, 模型需要检测出其中动作发生的开始时间、结束时间和动作类别. 而长尾时序动作检测则重点关注该任务的动作类别间不平衡问题. 本节将对本文提出的长尾时序动作检测方法进行详细介绍. 首先, 介绍长尾时序动作检测问题的相关定义和任务目标, 并展示所提出的基于半监督学习的长尾时序动作检测方法 WCReST 的整体框架. 之后, 具体描述该方法的类别重平衡采样策略. 最后, 说明如何计算被采样视频中每个伪标签动作片段的权重, 用以衡量其可靠性.

3.1 整体框架

本节首先介绍时序动作检测任务的基本算法. 如图 1 所示, 动作检测模型的任务目标是, 接受一个未剪辑视频片段作为输入, 输出视频中一个或多个动作的时间区间 (包括开始时间和结束时间) 以及动作的类别. 图 2 展示了动作检测模型的基本算法流程. 通常来说, 动作检测模型首先使用 I3D 模型[43] 等骨干网络来提取输入视频特征. 之后, 将提取得到的视频特征序列输入设计好的动作检测网络, 预测出多个动作框, 其中, 每个预测框包含一个动作片段的时间区间和动作类别概率. 最后, 将预测动作框和真实动作框进行对齐或匹配, 并计算模型的训练损失, 从而指导模型的迭代训练过程.

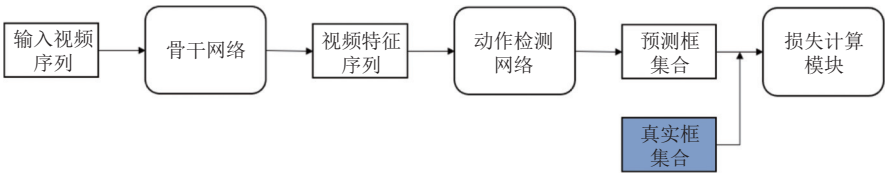


图 2 动作检测模型的基本算法流程

下面将对基于半监督学习的长尾时序动作检测问题进行系统地定义. 对于一个具有 K 个动作类别的时序动作检测任务, 半监督数据集可以被分为有标签数据集 D_0^L 和无标签数据集 D_0^U 两个部分.

与分类任务不同, 由于检测任务包括了定位和分类两个层次, 因此, 检测任务的样本 (视频) 和标签 (动作片段) 存在不对称性. 本文在动作片段级别上衡量标签的出现频率. 对于 K 个动作类别 $C=\{c_a\}_{a=1}^K$, 将有标签训练集 D_0^L 中动作类别 c_a 的动作片段标注数记为 N_a. 在不影响泛化性能的情况下, 假设 $C=\{c_a\}_{a=1}^K$ 中的动作类别按所包含的片段标注数量降序排序, 即 $N_1 \geqslant N_2 \geqslant \ldots \geqslant N_K$. 本文使用类别不平衡因子 $\eta=N_1/N_K$ 来衡量数据集的类别不平

衡性, η 越大代表数据集的不平衡性越严重.

对于无标签数据集 D_0^U, 本文假定 D_0^U 与有标签数据集 D_0^L 具有相近似的数据分布. 在检测任务中, 该假设包含两个方面: 一方面, 假定无标签数据集 D_0^U 中动作片段层面上的动作类别分布与有标签数据集 D_0^L 中相一致; 另一方面, 假设无标签数据集 D_0^U 中的每个视频样本都存在动作片段, 且每个视频样本所包含的平均动作片段数与有标签数据集 D_0^L 中相近似. 本文将有标签视频样本数占所有视频样本数的比例记为有标签比率 $\beta = N/(N+M)$, 其中, N 和 M 分别表示有标签数据集 D_0^L 和无标签数据集 D_0^U 中视频样本的数量.

给定数据具有长尾分布效应的有标签数据集 D_0^L 和无标签数据集 D_0^U, 半监督长尾时序动作检测的任务目标是, 学习到一个在类别平衡的测试标准下也能泛化性能良好的检测模型. 具体来说, 可以按照有标签数据集 D_0^L 中包含的动作片段数, 将动作类别划分为头部类别、中部类别和尾部类别, 并要求模型在这 3 种类别上都能获得良好的检测性能.

图 3 展示了本文为长尾时序动作检测任务所设计的加权类别重平衡自训练方法 (WCReST) 的整体框架流程图. 如图 3 所示, 训练数据集包含有标签数据集 D_0^L 和无标签数据集 D_0^U 两个部分, WCReST 方法基于这两部分数据集不断迭代训练模型, 并在第 n 次迭代后, 得到最终的时序动作检测模型 M_n.

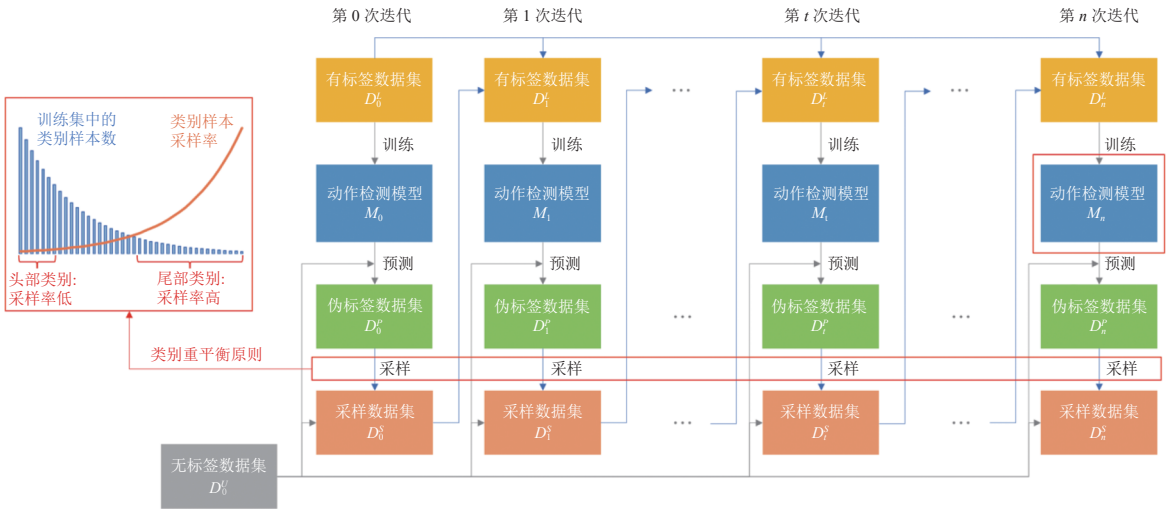


图 3 基于半监督学习的 WCReST 长尾时序动作检测方法流程图

而在第 t 轮迭代训练过程中, 本文提出的 WCReST 方法将完成以下几个步骤.

(1) 在有标签数据集 D_t^L 上训练得到基线时序动作检测模型 M_t.

(2) 使用模型 M_t 为无标签数据生成伪标签, 得到伪标签数据集 D_t^P.

(3) 基于类别重平衡采样策略, 从伪标签数据集 D_t^P 中采样带有伪标签的视频样本, 获得采样数据集 D_t^S.

(4) 将采样数据集 D_t^S 和初始有标签数据集 D_0^L 相结合, 得到下一轮用于训练的有标签数据集 $D_{t+1}^L = D_0^L \cup D_t^S$. 特别地, 被采样的视频样本中, 每个伪标签动作片段都被标记了可靠性权重, 这表示了它在下一轮迭代训练过程中的重要性.

3.2 类别重平衡采样策略

受文献 [40] 启发, 本文所提出的 WCReST 方法也基于类别重平衡原则来采样伪标签数据集中的视频, 更频繁地采样出现频率较低的动作类别样本.

具体来说, 初始有标签数据集 D_0^L 中包含 K 个动作类别 $C = \{c_a\}_{a=1}^K$, 且动作类别按所包含片段标注数降序排序, N_a 表示动作类别 c_a 的片段标注数. 在每轮迭代过程中, 本文希望按 D_0^L 中的类别分布来采样伪标签动作片段, 而不是将所有符合要求的伪标签片段都添加进有标签数据集. 具体来说, 在采样数据集 D_t^S 中, 为每个动作类别 c_a 直接

采样 $\hat{N}_a$ 个伪标签动作片段, $\hat{N}_a$ 的计算公式如下:

$$\begin{cases} \hat{N}_a = \mu_a \hat{N}'_a \\ \mu_a = \left(\dfrac{N_{K+1-a}}{N_1}\right)^{\alpha} \end{cases},$$

其中, $\hat{N}'_a$ 指动作类别 c_a 符合要求的所有伪标签片段数; μ_a 表示类别 c_a 的采样率; 采样参数 $\alpha \geqslant 0$ 是调节采样数据集 D_t^S 中标签分布的超参数, 它也决定了 D_t^S 的样本规模.

这种采样方法主要出于两点考量: 第一, 相关研究[1]发现, 在类别不平衡数据上训练得到的分类模型更倾向于将样本预测为头部类别, 因此往往在头部类别上获得较高的召回率, 而在尾部类别上具有较高的准确率, 所以尾部类别的伪标签样本具有相对更高的可靠性; 第二, 更频繁地采样尾部类别样本加入训练集中, 可以帮助改善初始有标签数据集中的类别不平衡现象, 从而使模型的训练更加均衡.

然而, 与分类任务不同, 时序动作检测中每个视频的伪标签并不是一个单独的标签, 而是由一组动作片段的伪标签组成. 同一视频中的伪标签片段具有不同的动作类别和置信度, 但又相互关联, 不可分割. 且动作片段的定位和分类任务往往也需要视频中背景信息和其他片段的辅助. 基于此, 本文提出了加权类别重平衡采样方法 WCReST, 以视频为单元进行采样, 并为被采样视频中的每个伪标签片段添加权重参数, 用以表示其在后续训练中的可靠性.

具体来说, 第 t 轮迭代得到的伪标签数据集被记为 D_t^P, 且有:

$$\begin{cases} D_t^P = \{\hat{\varphi}_{i,t}\}_{i=1}^{N_t^P} \\ \hat{\varphi}_{i,t} = (\hat{x}_{i,t}, \hat{s}_{i,t}, \hat{e}_{i,t}, \hat{c}_{i,t}, \hat{p}_{i,t}) \end{cases},$$

其中, N_t^P 表示伪标签片段的数量, $\hat{x}_{i,t}, \hat{s}_{i,t}, \hat{e}_{i,t}, \hat{c}_{i,t}$ 和 $\hat{p}_{i,t}$ 分别表示伪标签片段 $\hat{\varphi}_{i,t}$ 的输入视频样本、开始时间、结束时间、动作类别和预测置信度. 特别地, 预测置信度 $\hat{p}_{i,t}$ 由模型预测的动作片段分类得分映射得到, 映射函数采用验证集上将所有片段分类得分归一化至区间 $[0,1]$ 的映射函数. 另外, 伪标签数据集 D_t^P 已预先经过阈值筛选, 即每个伪标签片段的置信度 $\hat{p}_{i,t}$ 都高于阈值 t_{low}. 直观上来说, 阈值 t_{low} 代表了被采样的伪标签片段的置信度下限, 置信度低于 t_{low} 的伪标签片段被认为是噪声伪标签, 不会被采样 (包括直接采样和间接采样) 进采样数据集. 这是因为, 置信度过低的伪标签片段不仅无法为模型的迭代训练提供有价值信息, 甚至还有可能干扰模型的正常训练.

而第 t 轮训练的采样数据集被记为 D_t^S, 且有:

$$\begin{cases} D_t^S = \left\{\left(x_{q,t}, \hat{\Psi}_{q,t}\right)\right\}_{q=1}^{N_t^S} \\ \hat{\Psi}_{q,t} = \{\hat{\psi}_{q,j,t} = (\hat{s}_{q,j,t}, \hat{e}_{q,j,t}, \hat{c}_{q,j,t}, w_{q,j,t})\}_{j=1}^{\hat{m}_{q,t}} \end{cases},$$

其中, N_t^S 表示被采样视频的数目, $x_{q,t}$ 表示第 q 个被采样的视频, $\hat{\Psi}_{q,t}$ 表示该视频中 $\hat{m}_{q,t}$ 个伪标签片段的集合. $\hat{s}_{q,j,t}$, $\hat{e}_{q,j,t}, \hat{c}_{q,j,t}$ 分别表示视频 $x_{q,t}$ 中第 j 个伪标签片段 $\hat{\psi}_{q,j,t}$ 的开始时间、结束时间和动作类别, 而 $w_{q,j,t}$ 则是其可靠性权重参数, 在之后的迭代训练过程中, $w_{q,j,t}$ 被作为第 j 个伪标签片段的损失权重, 参与训练损失的计算. 加权类别重平衡采样策略的具体步骤见算法 1.

算法 1. 加权类别重平衡采样策略.

输入: $D_t^P = \{\hat{\varphi}_{i,t} = (\hat{x}_{i,t}, \hat{s}_{i,t}, \hat{e}_{i,t}, \hat{c}_{i,t}, \hat{p}_{i,t})\}_{i=1}^{N_t^P}$;

输出: $D_t^S = \left\{\left(x_{q,t}, \hat{\Psi}_{q,t}\right)\right\}_{q=1}^{N_t^S}$.

1. $\tilde{D}_t^P = \varnothing$
2. **for** $a = K$ to 1 **do**
3. $\quad D_{a,t}^P = \{\hat{\varphi} | \hat{\varphi} = (\hat{x}, \hat{s}, \hat{e}, \hat{c}, \hat{p}) \in D_t^P$ **and** $\hat{c} == a$ **and** $\hat{p} \geqslant t_{\text{high}}\}$
4. $\quad$ Sort $D_{a,t}^P$ in descending order by $\hat{p}$
5. $\quad \hat{N}'_a = \textbf{len}\left(D_{a,t}^P\right)$
6. $\quad \hat{N}_a = \mu_a \hat{N}'_a$

$D^P_{a,t} = D^P_{a,t}[:\hat{N}_a]$
$\tilde{D}^P_t = \tilde{D}^P_t \cup D^P_{a,t}$
end for
$D^S_t = \varnothing$
for $\hat{\varphi} = (\hat{x}, \hat{s}, \hat{e}, \hat{c}, \hat{p})$ in $\tilde{D}^P_t$ **do**
  flag = **false**
  for $(\hat{x}_s, \hat{\Psi})$ in D^S_t **do**
    if $\hat{x} == \hat{x}_s$ **then**
      for $\hat{\psi} = (\hat{s}', \hat{e}', \hat{c}', \hat{w}')$ in $\hat{\Psi}$ **do**
        if $\hat{s} == \hat{s}'$ and $\hat{e} == \hat{e}'$ and $\hat{c} == \hat{c}'$ **then**
          $\hat{w}' = 1$
          break
        end if
      end for
      flag = **true**
      break
    end if
  end for
  if *flag* == **false then**
    $\hat{\Psi} = \{(\hat{s}', \hat{e}', \hat{c}', \hat{p}' \times \mu_{\hat{c}'}) | \hat{\varphi}' = (\hat{x}', \hat{s}', \hat{e}', \hat{c}', \hat{p}') \in D^P_t$ **and** $\hat{x}' == \hat{x}$ **and** $\hat{\varphi}' \neq \hat{\varphi}\} \cup \{(\hat{s}, \hat{e}, \hat{c}, 1)\}$
    $D^S_t = D^S_t \cup \{\hat{x}, \hat{\Psi}\}$
  end if
end for

算法 1 展示了第 t 轮迭代过程中的加权类别重平衡采样流程, 该流程主要由两个步骤组成.

第 1 步, 基于类别重平衡原则, 采样得到以伪标签动作片段为单元的数据集 $\tilde{D}^P_t$. 对于动作类别 c_a , 首先按照置信度高于阈值 t_{high} 筛选伪标签动作片段, 记满足条件的伪标签片段数为 $\hat{N}'_a$; 然后根据类别重平衡原则, 计算该类别被采样的动作片段数 $\hat{N}_a$, 并采样置信度最高的前 $\hat{N}_a$ 个伪标签片段. 直观上来说, 阈值 t_{high} 代表了被直接采样的伪标签片段的置信度下限, 置信度高于阈值 t_{high} 的伪标签片段被认为具有较多有价值信息; 模型则基于类别重平衡原则, 对这些有价值的伪标签片段进行直接采样.

第 2 步, 根据 $\tilde{D}^P_t$, 生成以视频为单元的采样数据集 D^S_t . WCReST 方法将 $\tilde{D}^P_t$ 中所有伪标签片段所属的视频都加入采样数据集 D^S_t , 并为视频中的每个动作片段添加权重参数. 每个视频中的伪标签片段 $\hat{\varphi} = (\hat{x}, \hat{s}, \hat{e}, \hat{c}, \hat{p})$ 可以被分为两类.

1) 该片段被直接采样, 即 $\hat{\varphi} \in \tilde{D}^P_t$, 则将该片段的权重 $\hat{w}$ 置为 1.

2) 该片段被间接采样, 即 $\hat{\varphi} \notin \tilde{D}^P_t$, 则将其权重 $\hat{w}$ 置为该片段的预测置信度 $\hat{p}$ 和动作类别采样率 $\mu_{\hat{c}}$ 的乘积, 公式如下:

$$\hat{w} = \hat{p} \times \mu_{\hat{c}},$$

其中, 间接采样的动作片段指未被选中, 但是附带在被采样视频中的伪标签片段. 直觉上来说, 预测置信度与伪标签片段的可靠性成正比; 同时, 前文也说明了, 在长尾数据集上, 模型往往倾向于拟合头部类别, 所以出现频率越高的类别往往有着较低的分类准确率. 因此, WCReST 方法使用预测置信度和类别采样率的乘积来作为伪标签片段的可靠性权重.

3.3 加权损失计算方法

在第 3.2 节中提到, 在类别重平衡采样的过程中, 算法为被采样视频中的每一个伪标签动作片段都添加了权重参数, 该权重参数将被应用于模型迭代训练的损失函数计算过程中.

现有的单阶段时序动作检测模型主要有两种损失函数计算方式: 基于密集预测范式的动作检测模型[15–19,22]根据预测片段与真实片段的时序交并比值, 将预测出的所有动作片段划分为正例样本和负例样本, 并计算每个预测片段的损失; 基于集合预测范式的动作检测模型[20,21]则寻找预测片段集合与真实片段集合的最优二分匹配, 根据匹配结果将预测片段划分为正例样本和负例样本, 并计算损失.

这两种方式虽然各有不同, 但其本质都是基于真实片段为预测片段添加动作或背景标签. 具体来说, 对于采样数据集中的视频样本 $\left(x,\hat{\Psi}=\{\hat{\psi}_j=(\hat{s}_j,\hat{e}_j,\hat{c}_j,w_j)\}_{j=1}^{\hat{m}}\right)$, 算法根据 $\hat{\psi}$ 中的伪标签为模型预测得到的每个片段 y 生成标签 $\hat{y}=(\hat{s},\hat{e},\hat{c},\hat{w})$. 其中, 若片段 y 被划分为动作正例样本, 且相对应的伪标签片段为 $\hat{\psi}_j=(\hat{s}_j,\hat{e}_j,\hat{c}_j,w_j)$, 则标签 $\hat{y}=\hat{\psi}_j$; 若片段被划分为背景负例样本, 则标签 $\hat{\psi}_j=(\text{None},\text{None},\hat{c},1)$, $\hat{c}$ 是背景类别, 而权重参数被置为 1. 每个预测片段的损失函数计算公式如下:

$$L(y,\hat{y})=\hat{w}L'(y,\hat{y}),$$

其中, $L'(y,\hat{y})$ 表示所使用的基线动作检测算法的原有损失函数.

这种加权损失计算方式可以减轻被误分类为正例的噪声伪标签片段的影响, 伪标签片段的可靠性较低表示它更可能是噪声伪标签片段, 相应地, 其训练损失也显著小于被直接采样的伪标签片段. 与此同时, 该方法也可以避免动作片段被误分类为背景片段, 如果算法只采样高置信度的伪标签动作片段 (即只保留直接采样的伪标签片段), 迭代训练出的模型可能会倾向于将动作片段预测为背景, 导致模型性能的衰退. 此外, 研究表明, 时序动作检测算法在对预测片段进行划分时, 往往会出现负例样本远多于正例样本的情况, 尤其是使用二分匹配的基于集合预测范式的动作检测算法, WCReST 方法保留更多的正例动作伪标签片段, 也是出于这一点考量.

另一方面, 时序动作检测模型需要同时完成定位和分类任务, 因此其损失函数也往往需要考虑定位和分类两个方面. 现有的时序动作检测算法的损失函数大多由分类损失、定位损失、质量损失和其他损失组成. 由于算法无法保障所生成的伪标签片段的时序区间的精确性, 且模型在动作定位任务上的性能与动作类别的关联程度较小, 因此本文认为, 自训练方法对提升模型定位能力的贡献相对较小. 而质量损失通常用于评估预测框的完整度, 与动作类别无关, 因此受数据类别不平衡的影响也较小. 基于上述这两点, 在迭代训练的过程中, WCReST 方法进行了视频级别上的损失函数加权, 降低了模型在伪标签视频样本上的定位损失权重和质量损失权重.

4 实验数据集介绍

本文在 THUMOS14[1] 和 HACS Segments[2] 这两个时序动作检测数据集进行了实验, 并分别利用 THUMOS15 数据集[3]和 ActivityNet1.3 数据集[4] 构建相对应的无标签数据集. 此外, 本文还收集并整理了 Dance 数据集, 并在其上进行了实验. 本节将对这 3 个半监督时序动作检测数据集进行介绍, 并说明 Dance 数据集的构建思路和方法.

4.1 数据集介绍

THUMOS14 数据集[1]收集自 YouTube 网站, 包含 20 个体育动作类别. 其验证集和测试集分别包含 200 个和 212 个未经剪辑的长视频, 并分别标注了 3007 个和 3358 个动作片段. 为了方便与之前工作比较, 本文将 THUMOS14 数据集的验证集作为有标签训练集, 并在 THUMOS14 测试集上进行测试. 图 4 展示了 THUMOS14 有标签训练集中的动作类别分布, 其类别不平衡因子 $\eta=16.6$. 至于无标签数据集, THUMOS15 数据集[3]是对 THUMOS14 数据集的扩展, 动作类别与 THUMOS14 一致, 且具有与 THUMOS14 数据集相近似的数据分布. 由于 THUMOS15 数据集未公开其测试集的标注信息, 本文使用 THUMOS15 测试集中的全部 5613 个未剪辑长视频来作为 THUMOS14 数据集相应的无标签数据集. 值得一提的是, 这些无标签视频中包含了大量未知数量的背景视频样本 (即不含任何预定义动作片段), 一定程度上增加了半监督学习的难度.

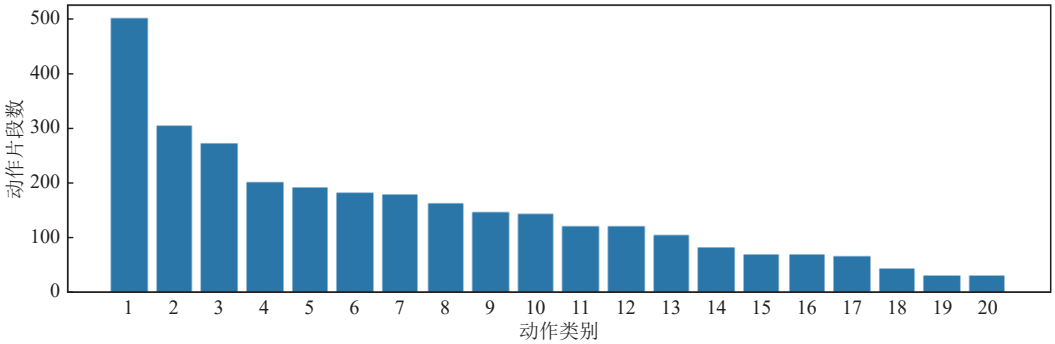


图 4　THUMOS14 验证集中的动作类别分布

HACS Segments 数据集[2]也来源于 YouTube 视频网站, 包含 200 个动作类别, 在 49 581 个未经剪辑的长视频标注了接近 14 万个动作片段, 训练集、验证集和测试集分别包含 37 613、5 981 和 5 987 个长视频. 由于 HACS Segments 数据集并未公开测试集的标注文件, 所以本文汇报在其验证集上的测试结果, 将其训练集作为有标签训练集. 图 5 展示了 HACS Segments 有标签训练集中的动作类别分布, 其类别不平衡因子 $\eta = 38.9$. 本文使用 HACS Segments 测试集和 ActivityNet1.3 数据集[4]中的视频来构建相应的无标签数据集. 本文对 ActivityNet1.3 中的样本进行了筛选, 去除了重复视频, 并尽可能使标签分布与 HACS Segments 训练集中近似. 最终, 无标签数据集中共包含 18 612 个未剪辑长视频, 数据集的有标签比率 β 约为 0.67.

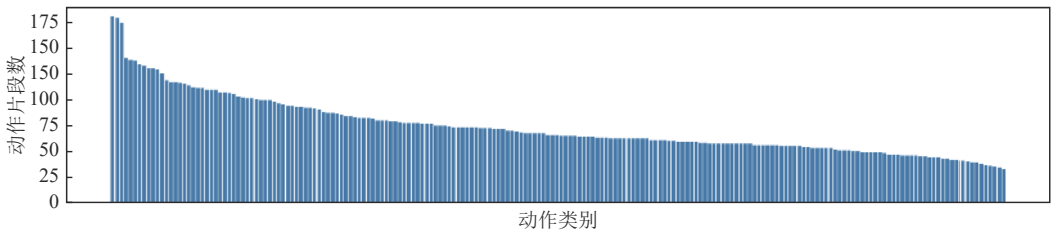


图 5　HACS Segments 训练集中的动作类别分布

Dance 数据集则是本文针对半监督长尾时序动作检测任务特点, 收集并整理的数据集. 该数据集收集自腾讯视频等视频网站的舞蹈分区, 包含 35 个动作类别、6 632 个有标签长视频和 13 264 个无标签长视频. 其中, 有标签视频按照 2∶1∶1 的比例划分训练集、验证集和测试集, 分别标注了 34 399、16 476 和 17 501 个动作片段. Dance 数据集的有标签比率 $\beta = 0.2$. 后文图 6 展示了 Dance 数据集中的动作类别分布情况, 可以看出, Dance 数据集保留了数据的长尾分布效应, 其训练集中类别的不平衡因子 η 为 84.2.

4.2　Dance 数据集构建

虽然 THUMOS14 数据集[1]和 HACS Segments 数据集[2]中也存在一定的类别分布不平衡性, 但如图 4 和图 5 所示, 由于预先经过了筛选, THUMOS14 验证集 (作为训练集使用) 和 HACS Segments 训练集中的长尾效应并不明显, 其不平衡因子 η 分别为 16.6 和 38.9.

因此, 为了验证本文所提出的 WCReST 方法在真实数据和现实应用场景下的性能表现, 本文针对舞蹈视频审核需求, 收集并整理了 Dance 数据集. 该数据集收集自腾讯视频、企鹅号等网络视频平台的舞蹈分区, 共包含 19 896 个未经剪辑的长视频, 视频的平均时长约为 197.6 s. 本文对这 19 896 个长视频进行了视频级别的动作类别标注, 共标注了 35 个舞蹈动作类别, 包括甩头、绕胸、摆胯等较为常见的舞蹈动作, 同时也关注了 M 腿、一字马等出现次数较少的舞蹈动作. 这些标签类别大多是预先定义好的舞蹈动作, 同时也在人工标注的过程中进行了补充.

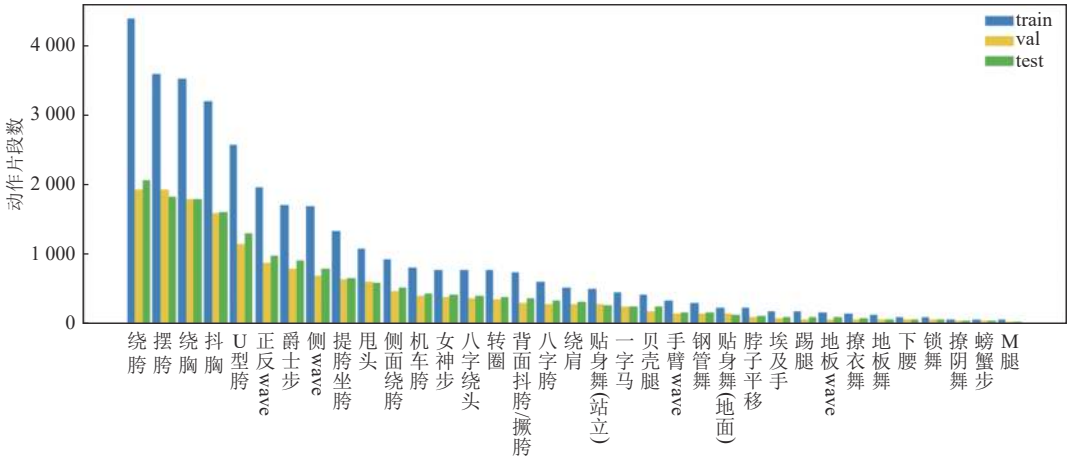


图 6　Dance 数据集中的动作类别分布

本文从 19 896 个长视频中随机抽取 6 632 个长视频, 对其进行时序动作检测标注, 构成 Dance 数据集的有标签部分. 相应地, 剩余的 13 264 个长视频则构成了 Dance 数据集的无标签部分. 本文将 Dance 数据集的有标签部分按 2:1:1 的比例划分为训练集、验证集和测试集, 因此, Dance 数据集的有标签比率 $\beta = 0.2$. 在抽取视频的过程中, 本文对每个动作类别按固定比例抽取样本, 从而尽可能保证无标签数据和有标签数据分布的一致性. 图 7 显示了在无标签数据集和有标签训练集上, 视频分类级别上的动作类别数据分布, 这展示了本文的有标签样本抽取原则. 在标注过程中, 为了保证标注的一致性和可靠性, 每个视频会被同一个标注者先后进行两次标注, 之后再经过另一位标注者的审核.

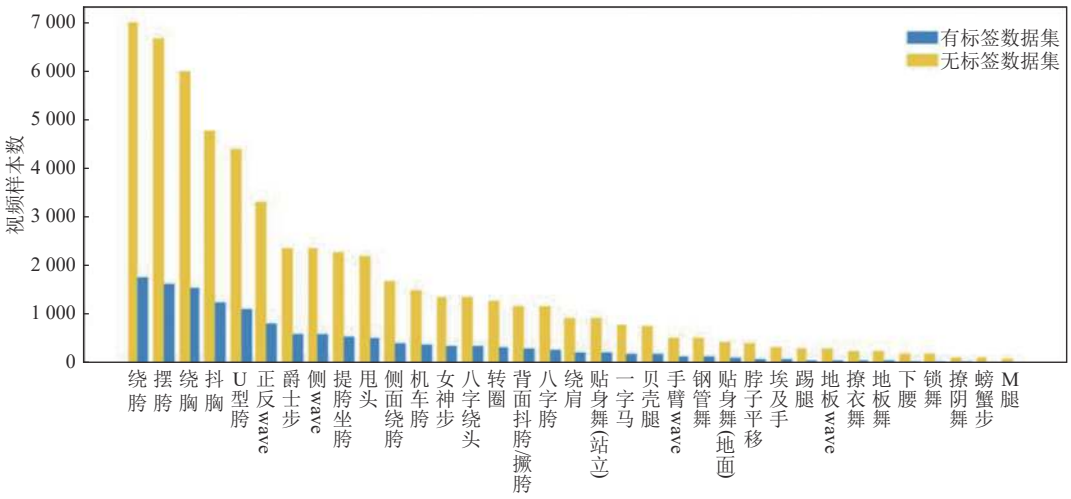


图 7　Dance 数据集在视频分类级别上的动作类别分布

Dance 数据集总计标注了 68 376 个动作片段, 其中, 训练集、验证集和测试集中分别包含了 34 399、16 476 和 17 501 个动作片段. 动作片段的平均时长约为 5.7 s. 图 6 展示了 Dance 数据集中动作类别的数据分布情况, 可以看出, 该数据集具有明显的长尾分布效应. 在 Dance 数据集的训练集中, 出现次数最多和最少的动作类别分别包含 4 378 个和 52 个动作片段, 数据集的不平衡因子 η 约为 84.2.

在舞蹈视频审核应用中, 需要检测的低俗违规视频大多包括 M 腿、贴身舞、撩阴舞等舞蹈动作. 从图 6 中可以看出, 这些动作的出现频率远低于正常舞蹈动作, 处于数据长尾分布的尾部区域. 因此, 模型对 Dance 数据集中

尾部类别的检测效果, 一定程度上决定了模型在视频审核实际应用中的性能表现.

5 实验结果分析

本节将通过实验来分析所提出的基于半监督学习的加权类别重平衡自训练方法 WCReST 在长尾时序动作检测任务上的效果. 首先, 本节会介绍实验中所使用的评价指标. 然后, 本节将描述实验的基线模型和训练参数细节. 接着, 本节将展示 WCReST 方法与基线模型的实验结果对比. 之后, 本节会通过详尽的消融实验来证明 WCReST 方法中各模块的有效性. 接下来, 本节将对比 WCReST 方法和其他深度长尾学习方法的实验结果. 最后, 本节将展示 WCReST 方法的可视化实验结果.

5.1 实验评价指标

本文所使用的实验指标为时序动作检测任务的常用评价指标 mAP, 即平均精度均值, 计算所有动作类别 P-R 曲线下面积的平均值. 此外, 本文还按照在训练集中动作片段的出现频率, 将数据集中的类别划分为头部类别 (frequent)、中部类别 (common) 和尾部类别 (rare), 并分别计算其 mAP 指标, 记为 $\mathrm{mAP_f}$、 $\mathrm{mAP_c}$ 和 $\mathrm{mAP_r}$, 用以衡量模型对不同类别的拟合效果. 表 1 展示了在所使用的 3 个数据集上, 这 3 种类别的动作片段出现次数所属的区间. 在 THUMOS14 数据集和 Dance 数据集上, 本文汇报模型在 tIoU 阈值集合 $\{0.3, 0.5, 0.7\}$ 上的 mAP 指标, 以及在 tIoU 阈值集合 $[0.3 : 0.1 : 0.7]$ 上的 average mAP 指标和 3 种类别集合的 average mAP 指标. 在 HACS Segments 数据集上, 本文汇报模型在 tIoU 阈值集合 $\{0.5, 0.75, 0.95\}$ 上的 mAP 指标, 以及在 tIoU 阈值集合 $[0.5 : 0.05 : 0.95]$ 上的 average mAP 指标和 3 种类别集合上的 average mAP 指标. 为了方便表示, 本文用 mAP_θ 表示 $\mathrm{tIoU} = \theta$ 时的 mAP 值, 用 mAP 表示 average mAP 值.

表 1 3 个数据集上, 头部、中部、尾部类别动作片段的出现次数区间

数据集	头部类别	中部类别	尾部类别
THUMOS14	>300	100–300	⩽100
HACS Segments	>1 200	200–1 200	⩽200
Dance	>2 000	200–2 000	⩽200

5.2 基线模型与参数设置

本文选取了 3 个有代表性的单阶段时序动作检测方法来作为对比实验的基线模型, 分别是基于密集预测范式的 AFSD[19] 和 ActionFormer[22], 以及基于集合预测范式的 TadTR[20]. 其中, ActionFormer 采用 Transformer 编码器, 并取得了目前最优的检测结果. AFSD[19]、ActionFormer[22] 和 TadTR[20] 都以视频为输入, 输出一组动作提名, 每个动作提名都包括一个动作片段的开始时间、结束时间和动作类别置信度.

AFSD[19] 模型首先使用骨干卷积神经网络来提取视频特征, 并转为 1D 特征金字塔; 然后, 对每个金字塔层, AFSD[19] 对每个时间点生成粗糙动作片段提名; 最后, AFSD 提取每个粗糙动作提名的边界显著性特征, 并基于该特征优化动作提名的时序边界, 并预测其动作类别.

ActionFormer[22] 模型同样使用骨干卷积神经网络来提取视频特征, 并使用编码器和解码器来预测每个时间点附近的动作片段提名. 在编码器中, ActionFormer[22] 使用带局部自注意力窗口的 Transformer 编码器来编码特征, 并穿插降采样层来关注不同尺度的动作; 在解码器中, ActionFormer[22] 使用 1D 卷积网络来解码特征, 输出每个时间点对应的动作提名信息.

TadTR[20] 模型也首先使用骨干卷积神经网络来提取视频特征; 然后, TadTR[20] 使用 Transformer 编码器来对特征进行编码; 最后, TadTR[20] 使用带动作查询的 Transformer 解码器来并行解码特征, 生成动作片段提名集合.

下面将详细描述本文的实验参数设置. 在 THUMOS14 数据集和 HACS Segments 数据集上, 本文使用与基线模型一致的特征提取方式和训练参数. 而在 Dance 数据集上, 本文使用 Kinetics 数据集上预训练的 I3D 模型[43] 来提取输入视频的双流特征. 每轮迭代训练中, 3 个基线模型的实现细节如下.

- AFSD 模型的视频帧采样 FPS 为 10, 滑动窗口大小为 256, 训练和测试时相邻窗口重叠帧数分别为 32 和

128. 定位损失权重 γ 设置为 5, 训练 epoch 数不超过 40, 测试时 Soft-NMS 操作的 tIoU 阈值设置为 0.5. 其余参数与原论文[19]中保持一致.

● ActionFormer 模型的特征时序下采样步长为 4, 训练时采用长度为 2 304 的可变输入特征片段. 局部自监督窗口大小设置为 16, 不添加位置编码. 模型训练的最大 epoch 数设置为 50, 线性预热 5 个 epoch, 学习率、小批量大小和权重衰减率分别设置为 0.000 1、2 和 0.000 1. 其余参数与原论文[22]中保持一致.

● TadTR 模型的特征时序下采样步长为 8, 滑动窗口大小为 128, 训练和测试时的窗口步长分别为 64 和 96, 动作查询数设置为 40. 训练的最大 epoch 数设置为 30, 并在 25 个 epoch 后降低学习率. 其余参数与原论文[20]中保持一致.

这一段将介绍与 WCReST 方法相关的参数设置. 模型在 THUMOS14、HACS Segments 和 Dance 数据集上的最大训练迭代轮数 n 分别为 6、8 和 10, 采样参数 α 被分别设置为 1、0.6 和 0.4, 置信度阈值 t_{high} 被分别设置为 0.95, 0.9 和 0.85, t_{low} 被设置为 0.5. 迭代训练时, 伪标签样本上的回归损失和质量损失被降低为原来的 1/2 .

5.3 与基线模型对比

本文在 THUMOS14、HACS Segments 和 Dance 这 3 个数据集上, 将所提出的 WCReST 方法和基线模型的实验结果进行了对比. 表 2–表 4 分别展示了在这 3 个数据集上的实验结果, 其中加粗部分表示同等实验条件下性能最好的模型指标结果.

表 2 在 THUMOS14 数据集上, WCReST 方法和基线模型的实验结果对比 (%)

方法	$\text{mAP}_{0.3}$	$\text{mAP}_{0.5}$	$\text{mAP}_{0.7}$	mAP	mAP_{f}	mAP_{c}	mAP_{r}
AFSD[19]	67.3	55.5	31.1	52.0	**72.2**	59.0	35.3
w/ WCReST	**67.8**	**56.8**	**32.1**	**52.9**	71.8	**59.8**	**36.7**
ActionFormer[22]	82.1	71.0	43.9	66.8	**83.0**	72.3	57.4
w/ WCReST	**82.8**	**72.6**	**44.7**	**68.8**	82.2	**73.0**	**58.4**
TadTR[20]	74.8	60.1	32.8	56.7	67.0	63.6	43.0
w/ WCReST	**75.3**	**61.2**	**33.2**	**58.1**	**67.7**	**64.6**	**45.1**

表 3 在 HACS Segments 数据集上, WCReST 方法和基线模型的实验结果对比 (%)

方法	$\text{mAP}_{0.5}$	$\text{mAP}_{0.75}$	$\text{mAP}_{0.95}$	mAP	mAP_{f}	mAP_{c}	mAP_{r}
TadTR[20]	47.14	32.11	10.94	32.09	—	—	—
TadTR*	47.06	32.35	10.99	31.93	38.61	32.28	22.03
w/ WCReST	**49.84**	**34.69**	**12.60**	**33.84**	**39.73**	**34.06**	**26.31**

注: *表示本文的复现结果

表 4 在 Dance 数据集上, WCReST 方法和基线模型的实验结果对比 (%)

方法	$\text{mAP}_{0.3}$	$\text{mAP}_{0.5}$	$\text{mAP}_{0.7}$	mAP	mAP_{f}	mAP_{c}	mAP_{r}
AFSD	66.32	48.71	25.76	45.87	72.43	50.52	23.28
w/ WCReST	**72.65**	**52.08**	**30.10**	**49.34**	**73.94**	**52.97**	**29.77**
ActionFormer	67.11	58.32	29.61	52.50	80.02	57.37	28.99
w/ WCReST	**71.28**	**63.82**	**37.61**	**55.31**	**80.88**	**59.35**	**34.45**
TadTR[20]	62.55	46.59	20.91	43.65	71.07	48.16	20.93
w/ WCReST	**64.54**	**50.29**	**24.16**	**46.22**	**71.95**	**50.41**	**24.96**

可以看出, WCReST 方法可以轻松应用于不同的基线模型. 在 HACS 数据集和 Dance 数据集上, WCReST 方法可以稳定提升基线模型的性能, 且对尾部类别的指标提升尤为明显. 这说明 WCReST 方法可以充分利用无标签数据中的有效信息, 提高模型对动作类别 (尤其是尾部类别) 的拟合能力. 而 THUMOS14 数据集的长尾效应并不明显, 且其无标签数据集包含了大量无动作片段的背景视频. 为了避免背景视频样本的干扰, 本文在 THUMOS14

数据集上设置了较高的置信度阈值, 因此每轮迭代采样的伪标签视频样本较少. 尽管如此, WCReST 方法仍能实现对基线模型性能的提升, 这证明了该方法对不同数据集的鲁棒性.

5.4 消融实验

这一部分将通过充分的消融实验, 来证明本文所提出的 WCReST 方法中每个部分的有效性. 本文有针对性地在 THUMOS14、HACS Segments 和 Dance 这 3 个数据集上都进行了实验, 下面对每组消融实验的实验结果进行展示和分析.

5.4.1 采样参数影响分析实验

图 8 展示了在 Dance 数据集上, 不同的采样参数 α 对模型的检测性能的影响, 图 8(a)–图 8(c) 分别使用 AFSD、ActionFormer 和 TadTR 作为基线模型. 其中, 横坐标表示采样参数 α 的取值, 纵坐标表示模型的 average mAP 指标与基线模型之间的差值, frequent、common、rare 和 all 分别表示头部类别、中部类别、尾部类别和全部类别集合. 可以看出, 随着 α 取值的变大, 模型对头部类别的拟合能力逐渐下降, 而对尾部类别的拟合能力逐渐提升. 这是因为 α 取值越大, 头部和中部类别的采样率就越低, 采样时的类别重平衡力度就越高. 而当 α 值提高到一定程度时, WCReST 方法对模型检测尾部类别能力的提升趋于饱和, 但仍会轻微降低对头部类别的检测能力, 这就带来了模型整体 mAP 指标的拐点. 在 Dance 数据集上, 该拐点约为 $\alpha = 0.4$.

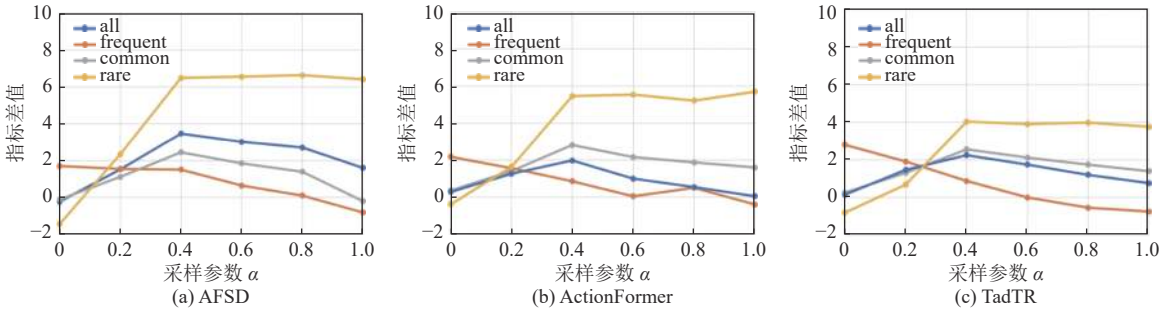


图 8 在 Dance 数据集上, 采样参数 α 的消融对比实验结果可视化

特别地, 当 $\alpha = 0$ 时, 所有类别的采样率都为 1, 模型退化为朴素采样策略, 即采样所有符合条件的伪标签动作片段, 表 5 展示了朴素采样策略和本文所提出的加权类别重平衡采样策略的实验结果对比. 实验结果表明, 虽然相较于朴素采样策略, WCReST 方法牺牲了一部分对头部类别检测能力的提升, 但它能够大幅度改善模型在尾部类别上的性能表现, 因此能够提升模型的整体性能.

表 5 在 Dance 数据集上不同采样策略的对比实验结果 (%)

基线模型	采样策略	mAP	mAP_f	mAP_c	mAP_r
AFSD	朴素采样策略	45.64	**74.13**	50.41	21.84
	WCReST	**49.34**	73.94	**52.97**	**29.77**
ActionFormer	朴素采样策略	52.87	**82.2**	57.66	28.61
	WCReST	**55.31**	80.88	**59.35**	**34.45**
TadTR	朴素采样策略	43.91	**73.88**	48.31	20.12
	WCReST	**46.22**	71.95	**50.41**	**24.96**

5.4.2 加权损失影响分析实验

本文所提出的加权损失主要有两个方面, 一方面是伪标签片段级别的权重计算, 为被采样视频中的每个伪标签动作片段添加权重, 用以表示其可靠性; 另一方面是视频样本级别的权重计算, 降低被采样的视频在训练时的定位损失和质量损失.

表 6 展示了对伪标签片段加权方法的消融对比实验结果, 其中第 2 列表示被间接采样的片段的权重 $\hat{w}$. $\hat{w} = 1$

表示将所有间接采样片段都作为直接采样片段处理; $\hat{w}=0$ 时则表示忽略所有间接采样的伪标签片段. 实验结果表明, 直接采样或者忽视所有被间接采样的伪标签片段都会造成模型性能的大幅度降低, 这是因为前者会受到大量被误分类为正例的背景噪声伪标签片段的干扰, 而后者会加剧视频样本中动作背景不平衡问题, 使模型倾向于将检测出的片段分类为背景. 此外, 在使用基线模型 ActionFormer 时, 本文还尝试比较了只使用预测置信度 $\hat{p}$ 或只使用动作类别采样率 $\mu_{\hat{c}}$ 来计算权重 $\hat{w}$, 表 6 中的对比实验结果证明了本文所提出的加权方法的有效性.

表 6　在 Dance 数据集上, 不同伪标签片段加权方法的对比实验结果 (%)

基线模型	间接片段权重 $\hat{w}$	mAP	mAP_f	mAP_c	mAP_r
AFSD	$\hat{w}=0$	40.38	65.78	44.03	20.37
	$\hat{w}=1$	45.87	72.43	50.52	23.28
	$\hat{w}=\hat{p}\times\mu_{\hat{c}}$	**49.34**	**73.94**	**52.97**	**29.77**
ActionFormer	$\hat{w}=0$	45.70	71.83	49.11	25.81
	$\hat{w}=1$	44.95	71.70	49.75	21.99
	$\hat{w}=\hat{p}$	53.24	79.83	57.28	31.85
	$\hat{w}=\mu_{\hat{c}}$	51.85	76.78	56.18	30.72
	$\hat{w}=\hat{p}\times\mu_{\hat{c}}$	**55.31**	**80.88**	**59.35**	**34.45**
TadTR	$\hat{w}=0$	39.65	65.3	43.55	19.04
	$\hat{w}=1$	38.07	61.97	43.37	15.51
	$\hat{w}=\hat{p}\times\mu_{\hat{c}}$	**46.22**	**71.95**	**50.41**	**24.96**

表 7 则对是否降低所有伪标签视频样本的定位损失权重和质量损失权重进行了讨论, 并展示其对比实验结果. 实验结果显示, 由于难以保证伪标签动作片段边界的准确性, 在不降低相关损失权重时, 会降低模型性能, average mAP 指标甚至会低于基线模型.

表 7　在 Dance 数据集上是否降低伪标签样本的定位和质量损失权重的对比实验结果 (%)

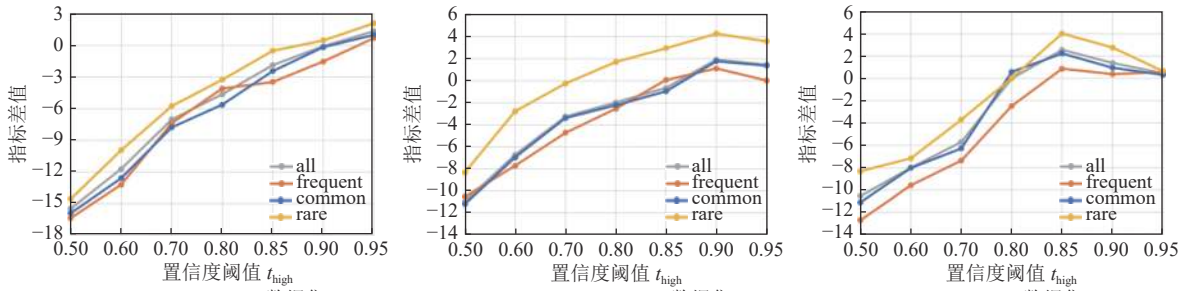
基线模型	是否降低损失权重	mAP	mAP_f	mAP_c	mAP_r
AFSD	不降低	46.68	70.68	50.85	26.34
	降低	**49.34**	**73.94**	**52.97**	**29.77**
ActionFormer	不降低	52.18	78.17	56.51	30.54
	降低	**55.31**	**80.88**	**59.35**	**34.45**
TadTR	不降低	43.16	69.0	47.62	21.33
	降低	**46.22**	**71.95**	**50.41**	**24.96**

5.4.3　置信度阈值影响分析实验

图 9 展示了在 THUMOS14、HACS Segments 和 Dance 数据集上, 不同置信度阈值 t_{high} 的对比实验可视化结果, 实验采用 TadTR 作为基线模型. 其中, 横坐标表示置信度阈值 t_{high} 的取值, 纵坐标和图例的相关含义与图 8 中相同.

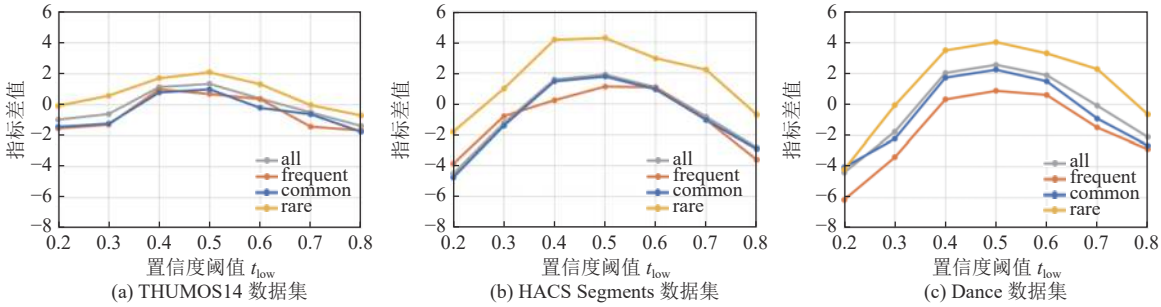
从图 9 中可以看出, 当 t_{high} 取值较小时, 模型的检测性能较差, 甚至远低于基线模型, 这是因为被采样的视频样本中存在较多被错误预测的伪标签片段, 干扰了模型的正常训练. 而当 t_{high} 取值过大时, 由于每次迭代采样的样本数量较少, 无法充分利用无标签数据中的信息, 因此也会造成 WCReST 方法性能的轻微衰退. 特别地, THUMOS14 数据集要求较高的置信度阈值 t_{high}, 这是因为其对应的无标签数据集中包含了大量背景视频样本, 需要通过高阈值来避免被误判段为正例的背景噪声样本的影响.

图 10 则展示了在 THUMOS14、HACS Segments 和 Dance 数据集上, 不同置信度阈值 t_{low} 的对比实验可视化结果, 实验同样采用 TadTR 作为基线模型. 其中, 横坐标表示置信度阈值 t_{low} 的取值, 纵坐标和图例的相关含义与图 9 和图 8 中相同. 实验结果显示, 过高和过低的置信度阈值 t_{low} 都会降低模型的性能表现. 这是因为当 t_{low} 过高时, 被间接采样的伪标签片段数量过少, 模型倾向于将片段预测为背景负例, 造成漏检测. 而当 t_{low} 过低时, 会带来大量带有错误伪标签的间接采样动作片段, 影响模型训练.



(a) THUMOS14 数据集 (b) HACS Segments 数据集 (c) Dance 数据集

图 9 在不同数据集上, 置信度阈值 t_{high} 的消融对比实验结果可视化, 使用 TadTR 作为基线模型



(a) THUMOS14 数据集 (b) HACS Segments 数据集 (c) Dance 数据集

图 10 在不同数据集上, 置信度阈值 t_{low} 的消融对比实验结果可视化, 使用 TadTR 作为基线模型

5.5 与其他深度长尾学习方法对比

本节将展示 WCReST 方法与其他深度长尾学习方法的实验结果对比. 由于在长尾时序动作检测任务上缺乏可比较的工作, 本文复现了部分其他领域的深度长尾学习方法, 包括针对长尾图像分类任务的 CReST[40] 和针对单阶段长尾目标检测任务的 EFL[29]. 其中, CReST[40] 基于半监督学习方法, 而 EFL[29] 则基于重加权方法. 后文表 8 展示了 CReST 方法、EFL 方法和本文所提出的 WCReST 方法在 Dance 数据集上的实验结果对比. 其中, CReST 方法等同于参数 $\hat{w}$ 取 0 的 WCReST 方法, 即忽略所有间接采样的伪标签动作片段, 实验结果与表 6 中 $\hat{w}=0$ 行相同. 可以看出, 本文所提出的 WCReST 方法具有领先的实验结果. 这是因为, WCReST 方法能够充分挖掘无标签数据中的有价值信息, 从而提升模型在有标签样本, 尤其是尾部动作类别样本上的性能表现. 另一方面, WCReST 方法针对时序动作检测任务的特点, 创新性地提出了伪标签动作片段的间接采样方法, 对无标签数据的利用更加高效、合理.

表 8 在 Dance 数据集上, 不同深度长尾学习方法的对比实验结果 (%)

基线模型	长尾学习方法	mAP	mAP_f	mAP_c	mAP_r
AFSD	CReST	40.38	65.78	44.03	20.37
	EFL	39.97	61.53	43.73	21.68
	WCReST	**49.34**	**73.94**	**52.97**	**29.77**
ActionFormer	CReST	45.70	71.83	49.11	25.81
	EFL	52.64	79.15	57.31	30.05
	WCReST	**55.31**	**80.88**	**59.35**	**34.45**
TadTR	CReST	39.65	65.3	43.55	19.04
	EFL	37.95	60.49	42.88	16.81
	WCReST	**46.22**	**71.95**	**50.41**	**24.96**

5.6 结果可视化

图 11 展示了 WCReST 方法对无标签视频数据进行加权类别重平衡采样的可视化流程, 基线模型选用 AFSD

模型. 输入视频为 Dance 数据集中的一段无标签视频样本. 该样本包含两个类别不同的动作片段, 标记在真实标签行中, 且用不同的颜色表示不同的动作类别. 预测伪标签行展示了预测得到的伪标签数据集, score 为每个伪标签动作片段的预测置信度; 加权伪标签行则展示了在采样数据集中, 该视频样本最终包含的伪标签片段集合, w 表示每个伪标签片段的权重. 在预测伪标签行和加权伪标签行中, 实线和虚线分别表示直接采样和间接采样的动作片段, 不同的颜色则代表不同的动作类别.

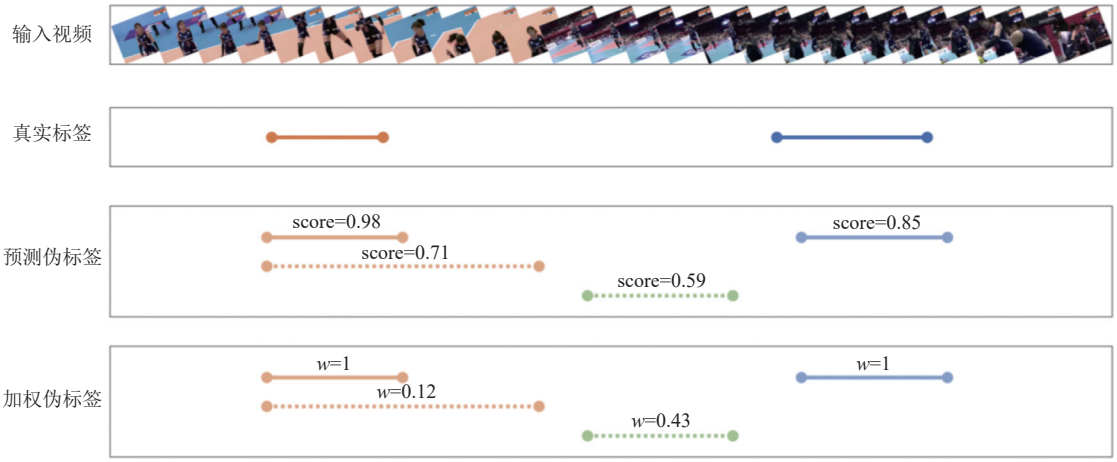


图 11 WCReST 方法采样流程的可视化展示图

可以看出, 本文所提出的 WCReST 方法能够充分挖掘无标签数据中的有效信息, 采样视频中的动作片段来进行自训练迭代; 同时, 通过降低间接采样伪标签的可靠性权重, WCReST 方法也能够在一定程度上减轻噪声伪标签的影响, 保证模型迭代训练的稳定性.

5.7 实验总结

本文使用多种基线模型, 分别在 THUMOS14 数据集、HACS Segments 数据集和 Dance 数据集上进行了实验. 实验结果表明, 本文提出的 WCReST 方法可以改善模型在尾部动作类别上的检测效果, 提升模型的整体性能表现, 并取得较为领先的实验结果. 同时, WCReST 方法可以轻松应用于多种基线模型, 具有一定的普适性和鲁棒性. 此外, 详尽的消融实验结果也证明了 WCReST 方法各个模块的有效性.

6 总结与展望

本文首次关注了时序动作检测任务中的长尾问题, 并基于半监督学习方法提出了一种加权类别重平衡自训练框架 WCReST. 通过在大量无标签数据上添加伪标签并迭代训练的方式, 来提高模型的动作检测性能. 在每轮迭代过程中, 本文根据训练集中动作类别的出现频率来采样伪标签样本, 从而重平衡有标签训练集中的动作类别分布, 提升模型对尾部类别的拟合能力. 针对动作检测任务特点, 本文将伪标签动作片段分为直接采样和间接采样这两部分, 并为间接采样的伪标签片段计算损失权重, 用以衡量其可靠性, 从而减少训练中错误伪标签的干扰. 此外, 本文还在视频级别上降低了伪标签样本的定位和质量损失, 从而保证模型迭代训练的稳定性. 特别地, 本文针对视频审核应用, 收集并整理了 Dance 时序动作检测数据集, 包含 35 个动作类别、6632 个有标签视频和 13264 个无标签视频, 并保留了数据中动作类别的显著长尾分布. 本文分别使用 AFSD、ActionFormer 和 TadTR 作为基线模型, 在 THUMOS14 数据集、HACS Segments 数据集和 Dance 数据集上进行了实验. 实验结果表明, 本文提出的 WCReST 方法可以提升基线模型的检测性能, 并且对尾部动作类别的性能提升尤为明显. 本文还通过详尽的消融实验, 证明了 WCReST 方法各个部分的有效性. 在之后的工作中, 可以对长尾时序动作检测任务进行进一步探索, 尝试使用更复杂的半监督学习方法和类别重平衡采样方法, 充分利用无标签数据中的有效信息, 并减少噪声伪标

签样本的影响.

References:

[1] Idrees H, Zamir AR, Jiang YG, Gorban A, Laptev I, Sukthankar P, Shah M. The THUMOS challenge on action recognition for videos “in the wild”. Computer Vision and Image Understanding, 2017, 155: 1–23. [doi: 10.1016/j.cviu.2016.10.018]

[2] Zhao H, Torralba A, Torresani L, Yan YC. HACS: Human action clips and segments dataset for recognition and temporal localization. In: Proc. of the 2019 IEEE/CVF Int’l Conf. on Computer Vision. Seoul: IEEE, 2019. 8667–8677. [doi: 10.1109/ICCV.2019.00876]

[3] Gorban A, Idrees H, Jiang YG, *et al*. THUMOS challenge 2015. 2015. http://www.thumos.info/

[4] Heilbron FC, Escorcia V, Ghanem B, Niebles JC. ActivityNet: A large-scale video benchmark for human activity understanding. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 961–970. [doi: 10.1109/CVPR.2015.7298698]

[5] Zeng RH, Huang WB, Gan C, Tan MK, Rong Y, Zhao PL, Huang JZ. Graph convolutional networks for temporal action localization. In: Proc. of the 2019 IEEE/CVF Int’l Conf. on Computer Vision. Seoul: IEEE, 2019. 7093–7102. [doi: 10.1109/ICCV.2019.00719]

[6] Lin TW, Zhao X, Su HS, Wang CJ, Yang M. BSN: Boundary sensitive network for temporal action proposal generation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–21. [doi: 10.1007/978-3-030-01225-0_1]

[7] Lin TW, Liu X, Li X, Ding ER, Wen SL. BMN: Boundary-matching network for temporal action proposal generation. In: Proc. of the 2019 IEEE/CVF Int’l Conf. on Computer Vision. Seoul: IEEE, 2019. 3888–3897. [doi: 10.1109/ICCV.2019.00399]

[8] Tan J, Tang JQ, Wang LM, Wu GS. Relaxed Transformer decoders for direct action proposal generation. In: Proc. of the 2021 IEEE/CVF Int’l Conf. on Computer Vision. Montreal: IEEE, 2021. 13506–13515. [doi: 10.1109/ICCV48922.2021.01327]

[9] Gao JY, Chen K, Nevatia R. CTAP: Complementary temporal action proposal generation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 70–85. [doi: 10.1007/978-3-030-01216-8_5]

[10] Gao JY, Yang ZH, Sun C, Chen K, Nevatia R. TURN TAP: Temporal unit regression network for temporal action proposals. In: Proc. of the 2017 IEEE Int’l Conf. on Computer Vision. Venice: IEEE, 2017. 3648–3656. [doi: 10.1109/ICCV.2017.392]

[11] Lin CM, Li J, Wang YB, Tai Y, Luo DH, Cui ZP, Wang CJ, Li JL, Huang FY, Ji RR. Fast learning of temporal action proposal via dense boundary generator. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 11499–11506. [doi: 10.1609/aaai.v34i07.6815]

[12] Qing ZW, Su HS, Gan WH, Wang DL, Wu W, Wang X, Qiao Y, Yan JJ, Gao CX, Sang N. Temporal context aggregation network for temporal action proposal refinement. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 485–494. [doi: 10.1109/CVPR46437.2021.00055]

[13] Xu MM, Zhao C, Rojas DS, Thabet A, Ghanem B. G-TAD: Sub-graph localization for temporal action detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10153–10162. [doi: 10.1109/CVPR42600.2020.01017]

[14] Shou Z, Wang DA, Chang SF. Temporal action localization in untrimmed videos via multi-stage CNNs. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1049–1058. [doi: 10.1109/CVPR.2016.119]

[15] Lin TW, Zhao X, Shou Z. Single shot temporal action detection. In: Proc. of the 25th ACM Int’l Conf. on Multimedia. Mountain: ACM, 2017. 988–996. [doi: 10.1145/3123266.3123343]

[16] Xu HJ, Das A, Saenko K. R-C3D: Region convolutional 3D network for temporal activity detection. In: Proc. of the 2017 IEEE Int’l Conf. on Computer Vision. Venice: IEEE, 2017. 5794–5803. [doi: 10.1109/ICCV.2017.617]

[17] Chao YW, Vijayanarasimhan S, Seybold B, Ross DA, Deng J, Sukthankar R. Rethinking the Faster R-CNN architecture for temporal action localization. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1130–1139. [doi: 10.1109/CVPR.2018.00124]

[18] Long FC, Yao T, Qiu ZF, Tian XM, Luo JB, Mei T. Gaussian temporal awareness networks for action localization. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 344–353. [doi: 10.1109/CVPR.2019.00043]

[19] Lin CM, Xu CM, Luo DH, Wang YB, Tai Y, Wang CJ, Li JL, Huang FY, Fu YW. Learning salient boundary feature for anchor-free temporal action localization. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3320–3329. [doi: 10.1109/CVPR46437.2021.00333]

[20] Liu XL, Wang QM, Hu Y, Tang X, Zhang SW, Bai S, Bai X. End-to-end temporal action detection with Transformer. IEEE Trans. on Image Processing, 2022, 31: 5427–5441. [doi: 10.1109/TIP.2022.3195321]

[21] Shi DF, Zhong YJ, Cao Q, Zhang J, Ma L, Li J, Tao DC. ReAct: Temporal action detection with relational queries. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 105–121. [doi: 10.1007/978-3-031-20080-9_7]

[22] Zhang CL, Wu JX, Li Y. ActionFormer: Localizing moments of actions with Transformers. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 492–510. [doi: 10.1007/978-3-031-19772-7_29]

[23] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[24] Pouyanfar S, Tao YD, Mohan A, Tian HM, Kaseb AS, Gauen K, Dailey R, Aghajanzadeh S, Lu YH, Chen SC, Shyu ML. Dynamic sampling in convolutional neural networks for imbalanced data classification. In: Proc. of the 2018 IEEE Conf. on Multimedia Information Processing and Retrieval. Miami: IEEE, 2018. 112–117. [doi: 10.1109/MIPR.2018.00027]

[25] He HB, Garcia EA. Learning from imbalanced data. IEEE Trans. on Knowledge and Data Engineering, 2009, 21(9): 1263–1284. [doi: 10.1109/TKDE.2008.239]

[26] Cui Y, Jia ML, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9260–9269. [doi: 10.1109/CVPR.2019.00949]

[27] Huang C, Li YN, Loy CC, Tang XO. Deep imbalanced learning for face recognition and attribute prediction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(11): 2781–2794. [doi: 10.1109/TPAMI.2019.2914680]

[28] Byrd J, Lipton Z. What is the effect of importance weighting in deep learning? In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 872–881.

[29] Li B, Yao YQ, Tan JR, Zhang G, Yu FW, Lu JW, Luo Y. Equalized focal loss for dense long-tailed object detection. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 6980–6989. [doi: 10.1109/CVPR52688.2022.00686]

[30] Tan JR, Lu X, Zhang G, Yin CQ, Li QQ. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1685–1694. [doi: 10.1109/CVPR46437.2021.00173]

[31] Liu ZW, Miao ZQ, Zhan XH, Wang JY, Gong BQ, Yu SX. Large-scale long-tailed recognition in an open world. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2532–2541. [doi: 10.1109/CVPR.2019.00264]

[32] Yin X, Yu X, Sohn K, Liu XM, Chandraker M. Feature transfer learning for face recognition with under-represented data. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5697–5706. [doi: 10.1109/CVPR.2019.00585]

[33] Huang C, Li YN, Loy CC, Tang XO. Learning deep representation for imbalanced classification. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5375–5384. [doi: 10.1109/CVPR.2016.580]

[34] Zhang X, Fang ZY, Wen YD, Li ZF, Qiao Y. Range loss for deep face recognition with long-tailed training data. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5419–5428. [doi: 10.1109/ICCV.2017.578]

[35] Kang BY, Xie SN, Rohrbach M, Yan ZC, Gordo A, Feng JS, Kalantidis Y. Decoupling representation and classifier for long-tailed recognition. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[36] Zhou BY, Cui Q, Wei XS, Chen ZM. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9716–9725. [doi: 10.1109/CVPR42600.2020.00974]

[37] Kang BY, Li Y, Xie S, Yuan ZH, Feng JS. Exploring balanced feature spaces for representation learning. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[38] Zhong ZS, Cui JQ, Liu S, Jia JY. Improving calibration for long-tailed recognition. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16484–16493. [doi: 10.1109/CVPR46437.2021.01622]

[39] Yang YZ, Xu Z. Rethinking the value of labels for improving class-imbalanced learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1618.

[40] Wei C, Sohn K, Mellina C, Yuille A, Yang F. CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 10852–10861. [doi: 10.1109/CVPR46437.2021.01071]

[41] Hyun M, Jeong J, Kwak N. Class-imbalanced semi-supervised learning. arXiv:2002.06815, 2020.

[42] Kim J, Hur Y, Park S, Yang E, Hwang SJ, Shin J. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1221.

[43] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733. [doi: 10.1109/CVPR.2017.502]

王雨虹(1998－), 女, 硕士, 主要研究领域为计算机视觉, 动作识别.

王利民(1988－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 深度学习.

武港山(1967－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 多媒体技术.