天津理工大学

硕士学位论文开题报告

学院名称 天津理工大学

专业名称 电子信息（计算机技术）

学生姓名 张 文

指导教师 高赞 职称 教授

入学年月 2023年8月

天津理工大学研究生院

填表日期 2024 年 11 月

说　　明

1．硕士学位论文作为综合衡量硕士研究生培养质量和学术水平的重要依据，应该是在所研究的学科领域或专门技术上做出一定贡献的科研成果。其选题应具有一定的先进性、创新性和适当的难度，并有足够的工作量。

2．各学院开题报告考核组原则上由具有硕士研究生导师资格的教师组成，一般不得少于 5 人，可聘请校外具有副高级职称及以上的同行专家。导师不得担任本人指导研究生的考核组组长。硕士研究生应向考核组作全面的开题报告，考核组对选题报告进行认真评审后，填写具体评审意见。

3．开题报告工作一般应于第三学期末前进行，并将硕士研究生的开题报告留存学院并归档保管。

论文题目	用于时序动作检测的多尺度特征提取与融合方法
开题日期	2024 年 11 月 20 日
阅读文献情况	阅读文献总量为 82 篇，其中阅读外文文献共计 73 篇。

开题报告内容

（以下各项标题不可随意更改，均可加页，参考文献需按照学位论文中对参考文献格式要求书写。）

（一）课题来源及研究的目的和意义

本人的课题题目是全监督时序动作检测，方向是与导师经过交流确定的研究方向。随着现在社交媒体以及抖音等短视频平台的崛起，越来越多的包含大量信息的视频数据出现在网上。在这庞大的短视频中，以人为中心的视频数据占据了绝大多数。如何使计算机实现以人为中心的视觉理解变得越来越重要，这可以为动作检索、目标跟踪、高亮检测、视频问答、监控人体异常行为等下游任务奠定基础[1]。因此，时序动作检测一直都是计算机视觉领域一个热门的研究课题。

时序动作检测(TAD)任务的目的是在视频中在准确地定位和识别所有的动作，包括动作的开始时间和结束时间，以及对应的动作类别标签[2]。时序动作检测又分为全监督(Fully supervised)方法和弱监督(Weakly supervised)方法，其中全监督方法需要对训练视频做大量的标注信息，需要提供完整帧级别的位置标签以及类别标签，而弱监督则不需要对训练视频做大量的标注，仅需提供视频级别的类别标签即可[3]。

目前针对该领域所提出的算法很多，但准确率并不高，数据也不够丰富，还远远达不到应用的程度，所以需要我们开发出新的更加精确的方法去解决该问题。

（二）国内外研究现状及分析

该方向主要用到的数据集包括 THUMOS14、ActivityNet v1.3、HACS、EPIC-Kitchens 100 等。THUMOS14 数据集[4]包含 413 个未剪辑的视频，其中验证集包含 200 个视频，测试集包含 213 个视频，具有 20 个动作类别，其动作类型主要是体育运动(如棒球投球、篮球扣篮、举重、跳水等)。THUMOS14 数据集中动作的最长时间大约是 10 秒，最短时间是 1 秒，平均时间在 2 到 3 秒之间。ActivityNet 是一个大型的视频理解数据集[5]，包含来自 YouTube 的 20k 个视频，

总时长超过了 600 小时，涵盖了 200 种不同类型的活动。该数据集被划分为三个子集：训练集包含 10024 个视频，验证集包含 4926 个视频，测试集包含 5044 个视频。EPIC-Kitchens 100 是最大的以自我为中心的动作数据集[6]，该数据集包含来自 700 个场景的 100 个小时的视频，主要是拍摄了不同厨房中的烹饪活动。该数据集视频的数量较少，但每个视频中动作实例的数量平均约为 128 个。HACS 数据集[7]包含 200 个动作类别，有 37613 个视频做为训练集以及 5981 个视频用于测试集。

在现有的研究中，通常采用不同 IoU 阈值下的 mAP 的平均值做为衡量算法性能的评价指标。例如在 THUMOS14 数据集上，采用 IoU 阈值为[0.3:0.1:0.7]的 mAP 的平均值做为衡量指标(即 IoU 阈值从 0.3 到 0.7，每隔 0.1 计算一次 mAP，将不同阈值下的 mAP 的平均值做为最终的性能指标，记作 mAP@[0.3:0.1:0.7])。在 ActivityNet v1.3 数据集与 HACS 数据集上，采用 mAP@[0.5:0.05:0.95]做为性能指标。在 EPIC-Kitchens 100 数据集上，采用 mAP@[0.1:0.1:0.5]做为性能指标。

早期该领域的解决方法主要是两阶段方法：时序动作提议生成阶段和动作分类阶段。首先通过单独的方法去生成一些视频片段做为动作提议，然后再对每个有效的动作提议进行分类和时间边界细化。R-C3D[8]使用一个 3D 卷积网络对视频流进行编码，生成可能包含动作片段的提议，随后对提议进行分类和微调。SSN[9]依赖 TAG 的策略来生成高质量的动作提议，并将每个动作实例分解为开始、过程和结束，此外还为每个阶段建立时间金字塔表示。BSN[10]预测每个时间点属于某个动作的开始概率、结束概率以及动作概率，并采用从局部到全局的方式生成高质量的动作提议。BMN[11]提出了一种端到端的动作提议生成方法：边界匹配网络，用来评估提议的置信度从而改进提议生成。G-TAD[12]是一种基于图卷积神经网络的方法，它将视频片段表示为图节点，来学习时间上下文信息与语义信息，以便生成高质量的动作提议。BSN++[13]是对 BSN 的改进，采用一个互补边界生成器进行更精确的时间边界预测。TSI[14]为了解决短动作检测率极低的问题，设计了一种尺度不变的损失函数，并利用全局上下文信息进对边界检测进行了优化。VSGN[15]设计了一个跨尺度的图网络，并采用边界采样来细化提议。

由于两阶段方法需要进行两个独立的阶段，这就增加了计算成本和复杂性，使整个系统变得更加繁琐，在实际生活中更难得到应用。于是，一些工作开始尝试在不使用动作建议的情况下，在一个阶段中完成动作的分类和定位，这简化了动作检测的过程，使其更快更高效。Lin 等人提出了第一个单阶段时序动作检测方法 SSAD[16]，该模型是一个全部由一维卷积所构成的网络，使用预训练的模型对视频进行处理，得到特征序列，再将特征序列做为 SSAD 模型的输

入，最终输出预测结果。A2Net[17]在传统的 anchor-based 方法的基础上引入了 anchor-free 的模式，融合了两种模式的优点，使其能够更好地处理不同长度的动作实例。随着 Transformer 在机器翻译以及目标检测领域所取得的巨大成功，一些工作开始在时序动作检测任务中引入注意力机制。例如，TadTR[18]使用类似 DETR 的基于 Transformer 的解码器，该方法通过使用一组可学习的 query，自适应地从视频中为每个查询提取时序上下文信息，并直接预测具有上下文的动作实例。它的核心是一个时序可变形的注意力模块，可以选择性地关注视频中的一组稀疏的关键片段。另一种基于 Transformer 的时序动作检测方法则是将自注意力机制及其变形体应用到特征编码器。例如，ActionFormer[19]使用多尺度 Transformer 将嵌入的特征编码为特征金字塔，通过一个轻量级的卷积解码器进行解码，最后通过一个分类头和一个边界回归头得到最终的输出结果。TriDet[20]分析了自注意力机制中所存在的秩损失与高计算复杂度的问题，提出了一个基于卷积的可缩放粒度感知层(SGP)来编码特征金字塔，并提出了一个“三叉戟头”，通过评估边界框周围的相对概率分布来建模动作边界。最近，也有大量的工作针对自注意力机制的缺陷进行不同的优化，也都取得了一些进展。例如 ADSFormer[21]提出了一个自适应的双选择性多头令牌混合器来自适应地调整头部和通道维度中特征的权重，并允许充分挖掘可区别性特征以提高 TAL 的性能。目前单阶段 TAD 方法中，DyFADet[22]算法的性能是最优的。该算法提出利用动态卷积来进行动态特征学习，从而提高学习到的特征的可区分性。

随着计算机硬件的发展，GPU 的性能得到了大幅度的提高，因此也有一些工作开始致力于研究端到端的 TAD 算法。AFSD[23]将原始帧做为输入，提出了一个基于显著性的细化模块，来细化边界特征。由于计算限制，AFSD 将输入的分辨率下采样到 96×96。TALLFormer[24]采用长短期记忆模块来处理输入，仅对短期记忆模块中的数据进行反向传播来节约内存，减少计算量。AdaTAD[25]引入了一个时序信息适配器来减少内存，成功将主干扩展到 10 亿个参数，将输入视频帧扩展到 1536 帧，从而获得高效的检测性能。

目前现有的挑战如下：(1) 目标边界的不确定性：再进行检测时，目标动作和背景的边界往往是难以确定的。(2) 动作的时间跨度较大：动作片段的长短不一，较长的动作片段超过了 200s，而较短的动作则大概在 1s 左右。(3) 计算资源需求高：时序动作检测需要处理大量的视频数据，端到端方法对计算资源的要求较高。

（三）研究内容及拟采取的研究方法、技术路线、研究的特色及创新点

由于计算资源有限，本文采用单阶段 TAD 来节省计算资源，采用卷积和注意力机制互补

的方法来解决注意力机制的秩损失问题。近年来，单阶段 TAD 一般由预训练的主干网络、视频特征编码器、解码器(动作分类头和边界回归头)组成。

我提出的方法是在视频特征编码器部分使用一个 Multi-scale Temporal Feature Layer，这个模块通过使用一个多头交叉注意力来对预训练的特征进行编码，以获取全局上下文信息，另外通过多个不同大小的卷积提取局部特征，将全局上下文信息和局部信息融合，使用残差连接保留原始信息，保证深层网络的可训练性。使用这个模块，可以避免纯基于 Transformer 结构的秩损失问题，也可以获取不同感受野下的特征，有利于检测不同时间长度的动作片段。在编码器部分使用 MaxPool 进行下采样操作，以提取特征金字塔进行后续的分类和回归操作。在获取到特征金字塔之后，为了更充分的融合各金字塔层的特征，借鉴 BiFPN[26]的思想，设计了一个多尺度检测头。实现了特征金字塔不同层之间的双向跨尺度连接，允许特征在不同层级之间通过自上而下和自下而上的路径进行更全面的信息传递和融合。

目前的实验取得了一定的效果，在 THUMOS14 数据集上使用预提取的 VideoMAEv2 特征进行实验，mAP@[0.3:0.1:0.7]指标比目前单阶段的 SOTA 提高了接近 1.3 个点（70.5→71.79）。在 ActivityNet v1.3 数据集上，mAP@[0.50:0.05:0.95]与 SOTA 相近（降低了 0.03%）。这证明目前的研究方向是没有问题的，目前的多尺度特征提取模块还比较简单，后续还需要继续改进该模块。

（四）研究进度安排，预期达到的目标

研究的整体安排如下：

2024/3—2024/8：查阅相关领域文献,确定研究方向；

2024/9—2025/1：继续阅读文献，并完成第一个工作相关的的实验任务；

2025/3—2025/9：做好调研工作，开始进行新的实验；

2025/9—2026/1：完成大论文的撰写以及答辩相关工作。

预期达到的目标：在上述工作的基础上，能够对模块的调整，使其能够学到更多的有用信息，提高动作定位的精度。

（五）主要参考文献

[1] 杨佳汇.面向视频数据的时序动作检测算法研究[D].北京邮电大学,2024.
[2] 贾兴伟.基于差分和全卷积注意力网络的时序动作检测方法研究[D].西安理工大

学,2023.
[3] 曹佳晨.不同学习范式下的时序动作检测算法研究[D].杭州电子科技大学,2022.
[4] Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah,M. The THUMOS challenge on action recognition for videos “in the wild”. Comput. Vis. and Image Under. 155, 2017.
[5] Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J. ActivityNet: A large-scale video benchmark for human activity understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 961–970 (2015).
[6] D. Damen et al., Rescaling egocentric vision, 2020, arXiv:2006.13256.
[7] Zhao, H., Torralba, A., Torresani, L., Yan, Z. Hacs: Human action clips and segments dataset for recognition and temporal localization. In: ICCV (2019).
[8] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In ICCV, 2017.
[9] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In ICCV, 2017.
[10] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN:Boundary sensitive network for temporal action proposal generation. In ECCV, 2018.
[11] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In ICCV,2019.
[12] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In CVPR, 2020..
[13] Haisheng Su, Weihao Gan, Wei Wu, Yu Qiao, and Junjie Yan. Bsn++: Complementary boundary regressor with scalebalanced relation modeling for temporal action proposal generation. arXiv 2020.
[14] Liu, S., Zhao, X., Su, H., Hu, Z. (2021). TSI: Temporal Scale Invariant Network for Action Proposal Generation. In: ACCV 2020.
[15] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In ICCV, 2021.
[16] Lin, T., Zhao, X., Shou, Z. Single shot temporal action detection. In: ACM MM,2017.
[17] Yang, L., Peng, H., Zhang, D., Fu, J., Han, J. Revisiting anchor mechanisms for temporal action localization. IEEE Trans. Image Process., 2020.
[18] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. IEEE Trans. Image Process., 2022.
[19] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of

actions with transformers. In Eur. Conf. Comput. Vis., 2022.

[20] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In CVPR, 2023.

[21] Qiang Li, Guang Zu, Hui Xu, Jun Kong, Yanni Zhang, Jianzhong Wang. An Adaptive Dual Selective Transformer for Temporal Action Localization. In IEEE Trans. Multim. 2024.

[22] Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, Fan Li. DyFADet: Dynamic Feature Aggregation for Temporal Action Detection. In ECCV 2024 .

[23] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In CVPR, 2021.

[24] Ke Ning, Lingxi Xie, Jianzhuang Liu, Fei Wu, and Qi Tian. 2021. Interaction-Integrated Network for Natural Language Moment Localization. IEEE Transactions on Image Processing, 2021.

[25] Shuming Liu, Chen-Lin Zhang, Chen Zhao, Bernard Ghanem. End-to-End Temporal Action Detection with 1B Parameters Across 1000 Frames. In CVPR 2024.

[26] Mingxing Tan, Ruoming Pang, Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. In CVPR 2020.

<table>
<tr><td rowspan="7">考核组成员</td><td colspan="2">姓名</td><td>职称</td><td>所在单位</td><td>签字</td></tr>
<tr><td>组长</td><td>温显斌</td><td>教授</td><td>天津理工大学计算机科学与工程学院</td><td></td></tr>
<tr><td rowspan="5">成员</td><td>高赞</td><td>教授</td><td>天津理工大学计算机科学与工程学院</td><td></td></tr>
<tr><td>袁立明</td><td>副教授</td><td>天津理工大学计算机科学与工程学院</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">考核组评审意见</td><td colspan="3">考核组评审内容：

论文选题是否具有前沿性和创新性
是 □ 否 □

研究方案是否具有可行性
是 □ 否 □

技术路线是否具有合理性
是 □ 否 □</td><td colspan="2">考核组评审结果：

通过 □ 不通过 □</td></tr>
<tr><td colspan="5">考核组组长签字： 年 月 日</td></tr>
<tr><td>学院意见</td><td colspan="5">主管领导签章： 公章 年 月 日</td></tr>
</table>

注：1. 导师不得担任本人指导研究生的考核组组长。 2. 本表一式两份，学院留存归档。