基于时空交叉感知的实时动作检测方法

柯 逍[1,2,3],缪 欣[1,2,3],郭文忠[1,2,3*]

(1. 福州大学计算机与大数据学院,福建福州 350116;2. 福建省网络计算与智能信息处理重点实验室(福州大学),福建福州 350116;3. 空间数据挖掘与信息共享教育部重点实验室,福建福州 350003)

摘 要: 时空动作检测依赖于视频空间信息与时间信息的学习. 目前,最先进的基于卷积神经网络(Convolutionsl Neural Networks,CNN)的动作检测器采用2D CNN或3D CNN架构,取得了显著的效果. 然而,由于网络结构的复杂性与时空信息感知的原因,这些方法通常采用非实时、离线的方式. 时空动作检测主要的挑战在于设计高效的检测网络架构,并能有效地感知融合时空特征. 考虑到上述问题,本文提出了一种基于时空交叉感知的实时动作检测方法. 该方法首先通过对输入视频进行乱序重排来增强时序信息,针对仅使用2D或3D骨干网络无法有效对时空特征进行建模,提出了基于时空交叉感知的多分支特征提取网络. 针对单一尺度时空特征描述性不足,提出一个多尺度注意力网络来学习长期的时间依赖和空间上下文信息. 针对时序和空间两种不同来源特征的融合,提出了一种新的运动显著性增强融合策略,对时空信息进行编码交叉映射,引导时序特征和空间特征之间的融合,突出更具辨别力的时空特征表示. 最后,基于帧级检测器结果在线计算动作关联性链接. 本文提出的方法在两个时空动作数据集UCF101-24和JHMDB-21上分别达到了84.71%和78.4%的准确率,优于现有最先进的方法,并达到73帧/秒的速度. 此外,针对JHMDB-21数据集存在高类间相似性与难样本数据易于混淆等问题,本文提出了基于动作表示的关键帧光流动作检测方法,避免了冗余光流的产生,进一步提升了动作检测准确率.

关键词: 实时动作检测;多尺度注意力;时空交叉感知

基金项目: 国家自然科学基金(No.61972097,No.U21A20472);国家重点研发计划(No.2021YFB3600503);福建省科技重大专项(No.2021HZ022007);福建省自然科学基金(No.2021J01612,No.2020J01494)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)02-0574-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220859

Real-Time Action Detection Based on Spatio-Temporal Interaction Perception

KE Xiao[1,2,3], MIAO Xin[1,2,3], GUO Wen-zhong[1,2,3*]

(1. *College of Computer and Data Science, Fuzhou University, Fuzhou, Fujian* 350116, *China*;
2. *Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou, Fujian* 350116, *China*;
3. *Key Laboratory of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou, Fujian* 350003, *China*)

Abstract: Spatiotemporal action detection requires incorporation of video spatial and temporal information. Current state-of-the-art approaches usually use a 2D CNN (Convolutionsl Neural Networks) or a 3D CNN architecture. However, due to the complexity of network structure and spatiotemporal information extraction, these methods are usually non-real-time and offline. To solve this problem, this paper proposes a real-time action detection method based on spatiotemporal interaction perception. First of all, the input video is rearranged out of order to enhance the temporal information. As 2D or 3D backbone networks cannot be used to model spatiotemporal features effectively, a multi-branch feature extraction network is proposed to extract features from different sources. And a multi-scale attention network is proposed to extract long-term time-dependent and spatial context information. Then, for the fusion of temporal and spatial features from two different sources, a new motion saliency enhancement fusion strategy is proposed, which guides the fusion between features by encoding temporal and spatial features to highlight more discriminative spatiotemporal features. Finally, action tube links

收稿日期:2022-07-20;修回日期:2023-02-28;责任编辑:朱梅玉
*通讯作者:郭文忠

are generated online based on the frame-level detector results. The proposed method achieves an accuracy of 84.71% and 78.4% on two spatiotemporal motion datasets UCF101-24 and JHMDB-21. And it provides a speed of 73 frames per second, which is superior to the state-of-the-art methods. In addition, for the problems of high inter-class similarity and easy confusion of difficult sample data in the JHMDB-21 dataset, this paper proposes an action detection method of key frame optical flow based on action representation, which avoids the generation of redundant optical flow and further improves the accuracy of action detection.

Key words: real-time action detection; multiscale attention; spatio-temporal interaction perception

Foundation Item(s): National Natural Science Foundation of China (No.61972097, No.U21A20472); National Key Research and Development Plan of China (No.2021YFB3600503); Major Science and Technology Project of Fujian Province (No.2021HZ022007); Natural Science Foundation of Fujian Province (No.2021J01612, No.2020J01494)

1　引言

时空动作检测是近年来的热点研究问题. 在无人驾驶、安全监控、交通运输、人机交互系统等领域,实时时空动作检测的应用越来越广泛. 时空动作检测的目的是对未裁剪视频中的动作进行分类,并定位视频中每个动作实例的开始帧和结束帧. 与视频动作识别类似,时空动作检测需要构建有效的运动信息对视频动作进行分类,而不同的是时空动作检测还需对动作目标进行检测以及定位,面临着更大的挑战.

随着动作识别的深入研究,Shao等人[1]根据时间粒度将复杂动作进一步划分为易于区分的子动作进行识别. 如图1所示,根据时间粒度的划分,跳远可以分为助跑、起跳、落地3个子动作. 这些子动作持续时间短,相应的样本数量也相对较少. 以往的动作识别方法[2]通过改变视频长度、随机剪辑等方法对其进行数据增强,并采用长短期建模[3~6]进行分类. 然而视频随机剪辑打乱视频原有序列,容易破坏相邻帧间的时间依赖性,丢失视频原有的语义内容. 为此,本文对动作进行分析,发现如跑步、引体向上、棒球击球等动作存在持续循环发生的现象,根据这种特性,本文提出了一种简单的数据增强方法,在保证视频语义信息以及时间依赖性不受破坏的情况下,增强了数据包含的时序信息.

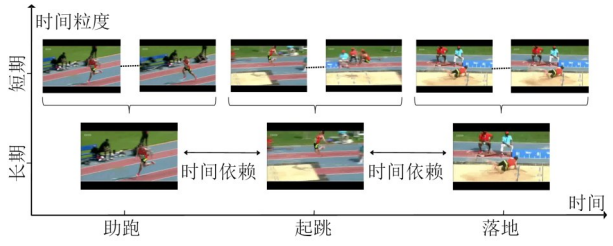


图1　动作的长短期粒度

为构建有效的运动信息,Feichtenhofer等人[7]研究发现在动作识别中光流信息能够有效地表示运动本身,通过简单的融合光流信息与空间信息,就能进一步改善动作识别精度. Zhao等人[8]研究发现光流信息能够有效区分如站立与坐下、抛球与接球等时间顺序性强、高类间相似性的动作. 如图2所示,光流信息能够有效区分此类易混淆动作,但因其计算十分耗时,难以应用于实时的检测. 最近的一些研究工作[9~11]试图找到光流的替代品,减少光流信息耗费的计算量以及存储空间,但这些替代品都存在一定的局限性,不能完整地表示运动信息. 因此,本文提出一种基于动作表示的关键帧光流数据输入方法,取代传统的光流数据输入,避免了光流噪声数据[12]的产生,有效地节约了光流信息的计算量和存储空间.

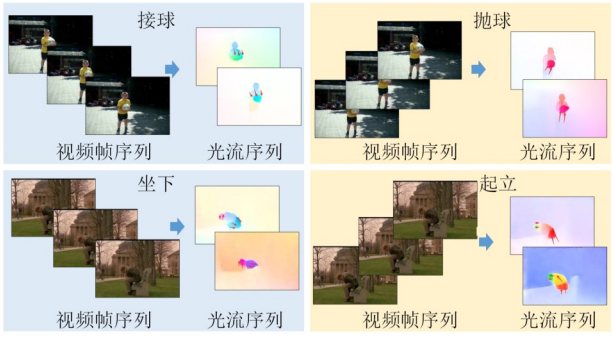


图2　光流的运动表示

时序信息与空间信息在时空动作检测中都起到十分重要的作用,如何有效地感知融合时序信息与空间信息也是时空动作检测的一个挑战. 近年来,一些对卷积神经网络进行改进的骨干网络,如P3D(Pseudo-3D)[13]卷积、R(2+1)D[14]卷积等在时空信息的提取上展现出明显的优势. 然而,这些方法大多数通过2D或3D卷积网络提取帧级或片段级特征,仅从单一尺度(即短期或长期)对时空特征进行描述,并直接将特征提取网络得到的时序特征与空间特征进行拼接,忽略了时序和空间特征数据来源不同,其特征中元素的关联关系也不同,不仅不能有效地融合时空特征,反而使得时序特征与空间特征互相排斥,混淆了视频图像中应当关注的运动区域. 基于这一想法,本文提出了基于多尺度的时空交叉感知注意力,对骨干网络提取的特征进行多尺度信息的增强,并引导时序特征和空间特征之间的融合,突出更具辨别力的时空特征表示.

综上所述,本文的主要贡献如下.

(1)针对时序动作的短期特征与长期特征进行分析,在保证视频语义信息以及时间依赖性不受破坏的前提下,通过对输入视频片段进行乱序重排,增强数据包含的时序信息.

(2)针对动作检测中易混淆动作引入光流信息进行难样本处理,不同于传统的光流数据输入,提出一种基于动作表示的关键帧光流数据输入方法,通过关键帧光流信息获取运动信息.与传统的数据输入相比,可以更清晰地获取运动信息并且有效地避免了噪声数据的产生,节约了光流信息的计算量和存储空间.

(3)提出基于多尺度的时空交叉感知注意力,主要包括多尺度注意力网络以及时空交叉变压器模块.不同于以往的多尺度注意力网络使用多级特征,本文仅使用骨干网络提取的特征通过不同扩张率的上下文注意力模块,扩大感受野,使其尽可能地覆盖所有尺度的对象,从而达到多尺度特征的效果.时空交叉变压器模块将多模态特征融合的思想引入时空特征融合中,通过对时空特征进行编码交叉映射,引导时序特征和空间特征之间的融合,突出更具辨别力的时空特征表示.

(4)对提出的各个模块进行了大量的消融实验,并以多个不同网络骨干作为特征提取网络在UCF101-24和JHMDB-21两个最常用数据集上验证了所提出方法的有效性.

2 相关工作

时空动作检测的发展与动作识别的相关研究息息相关.因此,在本节中,主要回顾当前动作识别、时空动作检测以及注意力机制的相关工作.

2.1 动作识别

与视频动作识别类似,时空动作检测需要构建有效的运动信息对视频动作进行分类.动作分类作为视频动作识别主要的研究方向之一,为动作检测奠定了基础.

现有的方法根据时序动作短期、长期时间依赖性的特点对视频的时序信息进行建模.Donahue等人[15]在2D模型基础上,通过长-短期记忆(Long Short-Term Memory,LSTM)提取时序特征.时间双线性网络[16](Temporal Bilinear Networks,TBN)、时间移位模块[17](Temporal Shift Module,TSM)、运动激励和聚合[18](Temporal Excitation and Aggregation,TEA)等都是2D网络的针对时序信息建模的变体.在不同的时空建模方法中,Simonyan等人[19]提出的双流网络达到了先进的性能.他的框架引入光流信息对动作进行时间建模,通过独立的并行网络从RGB和光流数据中提取特征进行时空信息建模.尽管这样的框架可以利用现有的2D CNN主干,但细粒度光流的提取非常昂贵.最近的研究工作[9~11]试图找到光流的替代品,如运动矢量、运动残差、Residual Frames等[11]方式来对动作进行建模,但都存在一定的局限性,不能完整地表示运动信息.

随着大规模动作数据集的出现(如kinetics),3D卷积神经网络表现出优越的时间建模能力,其产生的预训练模型更适合视频相关的任务.然而,与2D模型相比,3D模型固有的参数数量和计算成本更高.为了降低3D CNN的复杂性,P3D[13]卷积和R(2+1)D[14]卷积对每个3D解耦卷积转化为二维空间卷积和一维时间的卷积进行计算,从而进一步改进了3D卷积技术.

近年来,以Transformers为代表的算法在动作识别领域取得了良好的效果,Liu等人[20]提出一个端到端的时空动作检测框架TadTR(Temporal action detection with TRansformer),通过在视频中选取一些片段进行行动预测所需的时间背景信息,同时将所有动作都预测成一组标签和时间位置并列,简化动作管道的链接.Jacob等人[21]提出了一种新的基于Transformers的FAU(Facial Action Unit)相关网络,以捕获训练数据中广泛表达式的不同动作单元之间的关系.

2.2 时空动作检测

时空动作定位任务的目的是为在时间和空间上对视频动作实例进行分类,要求在动作发生的时间间隔内对动作进行正确的分类和准确的定位.其主要可以分为对运动目标的检测,以及动作的链接.现有的大多数时空动作检测方法[22~25]使用目标检测框架对视频帧的动作进行空间定位.Gkioxari等人[22]最先将R-CNN结构应用于动作边界框检测的每个帧并对动作进行分类,然后通过维特比算法将结果链接起来,设计了动作定位算法.然而,这种方法在链接过程中没有考虑动作的一致性,如果视频中存在多个目标动作,可能会导致性能降低.Saha等人[26]通过在链接后引入额外的标记操作解决了这个问题.为了更好地对检测的时间信息进行建模,Kalogeiton等人[27]提出了动作小管检测器.一些方法[28~31]在此基础上,通过小管道建议或建立单独训练链接网络生成管道建议进行分类.但额外的链接网络在增加准确率的同时也增加了模型的复杂性和运行时间.

2.3 注意力机制

注意力机制是增强卷积神经网络在视频中学习动作特征的有效方法[3~35].注意力建模是当前网络的重要组成部分,它可以帮助网络在不需要额外训练注释的情况下检测感兴趣的目标.Wang等人[36]首先将Non-Local注意力集成到神经网络中,以捕获视频帧中的长期依赖关系.Yue等人[37]进一步提出了按通道进行Non-Local注意力操作.Cao等人[38]提出了一个简化的Non-Local模块,该模块与压缩激励结构相结合,能够以较少的计算成本提高卷积神经网络的性能.Chen等

人[39]提出了全局推理单元,其中特征通道将在 Non-Local 模块之前压缩. EMANet(Expectation-Maximization Attention Networks)[40]模块和 TGM(Tensor Generation Module)[41]模块都考虑减少 Non-Local 的操作数量. EMANet[40]建议学习特征字典,而 TGM[41]模块则沿宽度、高度和通道方向分解其特征图. 考虑到视频区别于图像的时间属性,TEA[18]引入了运动激发(Motion Exctiation module, ME)和多时间聚集(Multiple Temporal Aggregation module, MTA)串联来捕捉短期和长期的时间变化. ACTION-net[33]提出了 STE 和 CE 模块,分别对时间和空间维度进行建模,解决了时空视角和时间维度上的通道之间的相互依赖性.

3　提出的方法

在本节中,首先对提出的动作定位方法进行概述,并对构建有效的动作表示、时空特征的增强以及融合等问题进行分析和讨论.

为寻找一种灵活的时空特征感知方法来表征时空动作,本文提出一种基于时空交叉感知的动作检测方法(Spatio-temporal Interactive Perceptual action Detection, SIPD). 图 3 展示了本文方法的网络架构,主要包括数据处理、特征提取网络以及多尺度时空交叉注意力模块. 首先,对输入的视频序列进行片段划分,进行乱序重排和关键帧提取,并计算基于动作表示关键帧的光流信息. 在特征提取网络中,将构造的新的视频序列、关键帧以及关键帧光流分别通过 3 个不同的骨干网络提取特征. 其中,3D 骨干网络提取时序信息,2D 骨干网络提取空间以及光流信息,具体骨干网络的选择将在实验部分进行探讨. 在多尺度时空交叉注意力中,通过多尺度注意力网络增强长期的时间依赖和空间上下文信息,而后对特征进行编码交叉映射,引导不同来源特征之间的融合,突出更具辨别力的时空特征表示. 最后,通过回归分类预测以获得更可靠的检测结果.

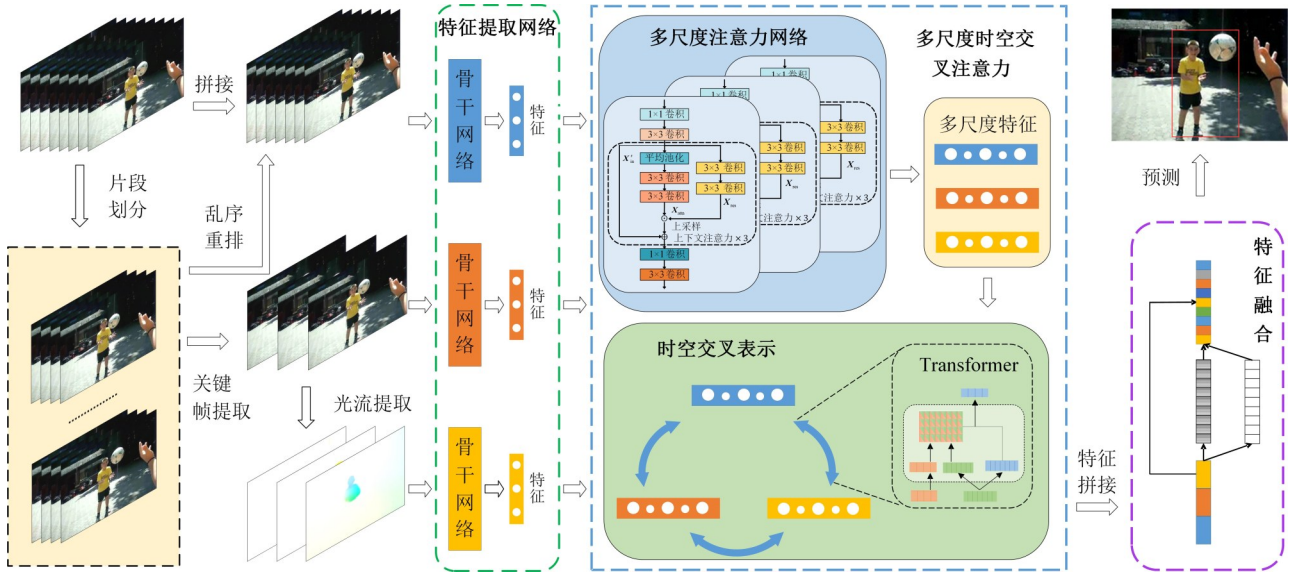


图 3　本文方法的模型网络架构

3.1　基于乱序重排的数据增强

受长短期建模和细粒度动作分类的启发,发现许多动作(如跑步、引体向上、棒球击球等)存在持续重复发生的现象. 对此,本文提出对动作进行乱序重排来增强动作的时序信息.

动作视频包含短期的动态信息以及动作的长期时间依赖性,提供了动作的时序结构. 本文的乱序重排数据增强方法通过对动作视频抽取连续的帧(片段)来构造视频片段集合,在小范围内对视频片段重新排序,构造新的视频序列,增加了数据的多样性. 由于是对视频片段集合中片段顺序的重排,所以并不会破坏视频片段中视频帧的短期时间依赖性,且小范围的顺序变化也保留了原视频动作序列的长期时间依赖性. 如先将一个输入视频划分为 n 个片段,构造视频片段集合 $\{s_1, s_2, \cdots, s_n\}$. 输入视频提供了动作的长期时序结构,而划分的视频片段,连续帧之间的时间依赖性提供了视频的短期时序结构. 而后,对构造的视频片段集合 $\{s_1, s_2, \cdots, s_n\}$ 进行重排序,n 个视频片段存在 $n!$ 种排列顺序,探索视频时间依赖性,在不破坏动作长期结构的前提下得到最佳的视频序列. 考虑到本文框架针对在线实时检测,输入视频帧序列长度有限,为确保视频片段中视频帧间的连续性以及视频动作的长期时序性,将构造的视频集合包含的片段数量 n 限制在 8 个以内. 计算方式如下:

$$S_{\text{new}}=\text{reorder}\left(\text{cut}\left(S,n\right)\right) \tag{1}$$

其中，S为输入视频序列，通过cut(·)构造函数得到视频片段元组$\{s_1,s_2,\cdots,s_n\}$，对视频片段元组进行reorder(·)重排序．将新的视频序列S_{new}与视频输入序列S连接，得到时序增强后的输入．

图4展示了对梳头动作片段进行乱序重排的过程，梳头动作可以进一步细分为抬手和下梳，这两个动作在长期时序上持续重复发生，我们将输入视频S进行抽样划分处理，得到n个等长的视频片段集合$\{s_1,s_2,\cdots,s_n\}$，每个片段s_i由等长的视频帧组成，每个视频片段可能为抬手或下梳．对这n个视频片段进行重排序，在不破坏其长期时序结构的前提下得到新的视频序列．

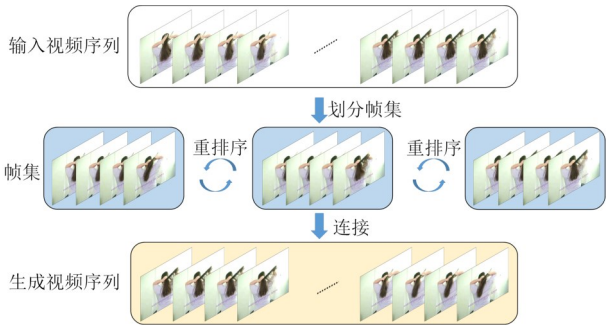


图4 乱序重排算法概述

3.2 基于动作表示的关键帧光流

如图2所示，光流信息通过计算像素点的位移和速度能有效区分如站立和坐下，抛球和接球等高类间相似性、易混淆的动作．然而，在连续的视频帧中，运动动作变化缓慢，全部光流提取存在大量的冗余信息，且随视频动作幅度变化，存在许多无用的噪声光流信息，对最后的运动特征表示也将产生影响．这一现象在相应的CNN特征图中得到了更多的体现．客观上，现有的光流信息计算方法也存在耗时、耗费大量存储空间等问题，难以实现实时场景下的动作检测．

因此，本文提出对基于动作表示的关键帧的进行光流信息的提取．本文提取关键帧的方法主要依据现实生活对动作过程的描述．如在生活中，为了解一个事件的大概情况，需要知道这个事件的起因、经过和结果，这也是对一个事件的简要描述．因此，本文通过将原始视频S进行cut(·)构造函数划分为n个子片段$\{s_1,s_2,\cdots,s_n\}$，然后通过select(·)函数随机抽取起始、中间、结尾片段(即s_1、$s_{n/2}$、s_n)中的帧作为关键帧$F=\{f_1,f_2,\cdots,f_m\}$．通过视频起始、中间、结尾的关键帧表示动作起始、中间以及结尾时的状态，通过关键帧表示动作的大概趋势．计算方式如下：

$$F=\text{select}\left(\text{cut}\left(S,n\right),m\right)=\{f_1,f_2,\cdots,f_m\} \tag{2}$$

如图5所示，对视频片段提取光流信息与对基于动作表示的关键帧提取光流信息的方法进行对比．可以看出，对整个视频片段进行光流提取，会出现如黄框所示的噪声片段．通过基于动作表示的关键帧光流信息的提取可以更清晰地获取运动信息并且有效地避免了噪声数据的产生，使得神经网络更好地关注到运动特征．

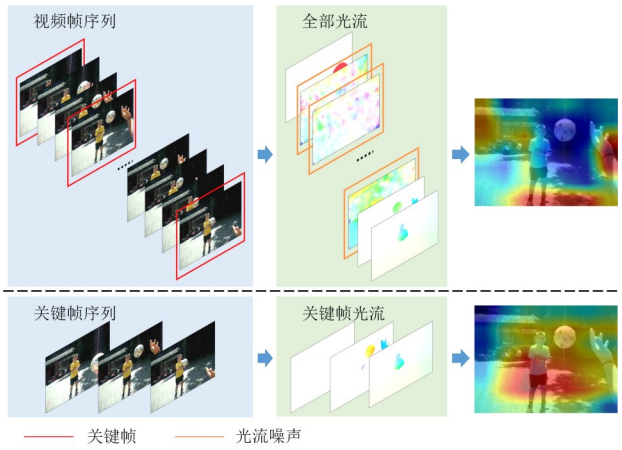


图5 基于动作表示的关键帧光流对比

3.3 时空特征增强与融合

本文提出基于多尺度的时空交叉特征融合注意力，主要包括多尺度注意力网络以及时空交叉变压器模块．

3.3.1 多尺度注意力网络

在目标检测中，多尺度特征融合应用十分广泛，现有的多尺度融合方法[42]主要通过不同网络层次具有不同大小的感受野，将多级特征进行连接作为下一层的输入．然而，Chen等人[43]研究表明特征金字塔最成功之处在于使用分治的策略优化问题而不是多尺度特征融合．

基于这一思想，本文设计了如图6(a)所示的多尺度特征增强模块，将特征提取网络提取的特征表示为$\boldsymbol{X}_{\text{in}}$．首先将$\boldsymbol{X}_{\text{in}}$输入卷积块得到新的特征表示$\boldsymbol{X}'_{\text{in}}=B\left(\boldsymbol{X}_{\text{in}}\right)$，其中，$B(\cdot)$包含一个1×1卷积层来降低信道维数，一个3×3卷积层来细化语义上下文．而后，通过堆叠多个不同卷积扩张率的上下文的注意力模块，使输出特征能够尽可能覆盖不同尺度上的所有对象．以图6(a)为例，我们堆叠了3个卷积扩张率分别为2、4、6的上下文注意力，逐步扩大感受野，生成具有多个感受野的输出特征．

其中，上下文注意力通过学习一种兼顾全局和局部的重加权机制，学习更具有代表性的语义特征．本文先通过卷积层生成局部信息特征映射，然后利用上下文特征对全局区域进行编码来确定特征映射的哪些区

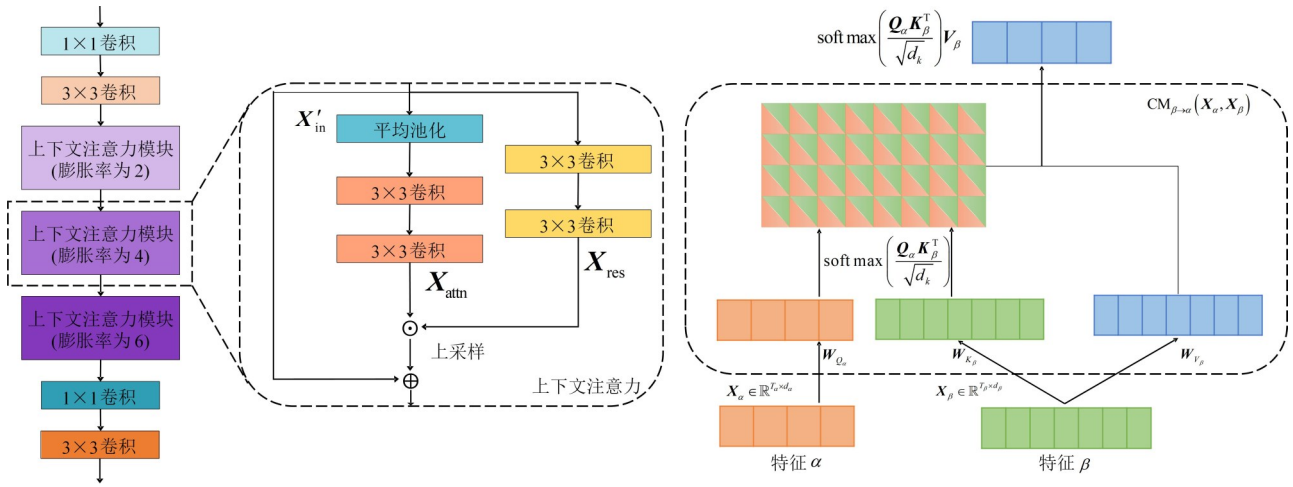
域应该被激活. 上下文注意力的计算方式为

$$\boldsymbol{X}_{\text{out}}=\boldsymbol{X}_{\text{attn}}{}^{*}\boldsymbol{X}_{\text{res}}+\boldsymbol{X}'_{\text{in}} \tag{3}$$

$$\boldsymbol{X}_{\text{attn}}=F_{\text{attn}}\left(\text{APool}\left(\boldsymbol{X}'_{\text{in}}\right);\theta,\Omega\right) \tag{4}$$

$$\boldsymbol{X}_{\text{res}}=F\left(\boldsymbol{X}'_{\text{in}};\theta,\Omega\right) \tag{5}$$

其中,$F(\cdot)$表示残差函数;APool$(\cdot)$表示平均池层;θ和Ω分别表示卷积块的结构. 我们使用APool$(\cdot)$来执行非完全压缩操作,然后对注意信道$\boldsymbol{X}_{\text{attn}}{}^{*}\boldsymbol{X}_{\text{res}}$的输出进行上采样,以匹配信道$\boldsymbol{X}'_{\text{in}}$的输出. 其中,$\boldsymbol{X}'_{\text{in}}\in\mathbb{R}^{T\times C\times H\times W}$和$\boldsymbol{X}_{\text{out}}\in\mathbb{R}^{T\times C\times H\times W}$为上下文注意力模块的输入和输出.



(*a*)多尺度注意力网络　　(*b*)特征潜在自适应表示

图6　多尺度时空交叉注意力结构概述

3.3.2　时空交叉变压器

为了能够更好地表示运动信息,需要对多尺度注意力网络提取的视频时序特征、空间特征以及光流特征进行融合. 基于这3个特征来源不同,特征元素间存在依赖的特性,本文将多模态特征融合的思想引入时空特征融合中,设计了时空交叉变压器,引导特征之间的融合,突出更具辨别力的时空特征表示. 首先压缩特征的空间维数,并对得到一维特征序列进行位置嵌入,保留序列的时间信息. 然后对不同来源特征进行交叉表示.

以时序特征和空间特征的交叉表示为例,通过解码器转换器,编码计算时序特征与空间特征的潜在自适应性. 将时序特征α和空间特征β的序列表示为$\boldsymbol{X}_\alpha\in\mathbb{R}^{L_\alpha\times d_\alpha}$和$\boldsymbol{X}_\beta\in\mathbb{R}^{L_\beta\times d_\beta}$,其中,$L(\cdot)$和$d(\cdot)$分别表示序列长度和特征维数.

如图6(*b*)所示,从空间特征β到时序特征α的潜在适应表现为$\boldsymbol{Y}_\alpha=\text{CM}_{\beta\to\alpha}\left(\boldsymbol{X}_\alpha,\boldsymbol{X}_\beta\right)\in\mathbb{R}^{L_\alpha\times d_v}$:

$$\boldsymbol{Y}_\alpha=\text{CM}_{\beta\to\alpha}\left(\boldsymbol{X}_\alpha,\boldsymbol{X}_\beta\right)=\text{softmax}\left(\frac{\boldsymbol{Q}_\alpha\boldsymbol{K}_\beta^{\text{T}}}{\sqrt{d_k}}\right)$$

$$\boldsymbol{V}_\beta=\text{softmax}\left(\frac{\boldsymbol{X}_\alpha\boldsymbol{W}_{Q_\alpha}\boldsymbol{W}_{K_\beta}^{\text{T}}\boldsymbol{X}_\beta^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{X}_\beta\boldsymbol{W}_{V_\beta} \tag{6}$$

定义$\boldsymbol{Q}_\alpha=\boldsymbol{X}_\alpha\boldsymbol{W}_{Q_\alpha}$,$\boldsymbol{K}_\beta=\boldsymbol{X}_\beta\boldsymbol{W}_{K_\beta}$,$\boldsymbol{V}_\beta=\boldsymbol{X}_\beta\boldsymbol{W}_{V_\beta}$. 其中,$\boldsymbol{W}_{Q_\alpha}\in\mathbb{R}^{d_\alpha\times d_k}$,$\boldsymbol{W}_{K_\beta}\in\mathbb{R}^{d_\beta\times d_k}$,$\boldsymbol{W}_{V_\beta}\in\mathbb{R}^{d_\beta\times d_v}$. 通过softmax计算$\boldsymbol{X}_\alpha,\boldsymbol{X}_\beta$对应的注意力得分矩阵(softmax$(\cdot)\in\mathbb{R}^{L_\alpha\times L_\beta}$),矩阵中第$(i,j)$项计算特征$\alpha$的第$i$个位置对特征$\beta$的第$j$个位置的注意权重. 因此,$\boldsymbol{Y}_\alpha$的第$i$个位置是$\boldsymbol{V}_\beta$的加权汇总,权重由softmax$(\cdot)$中的第$i$行确定.

然后,将特征潜在自适应表示嵌入Transformer中,通过多头注意模块、前馈网络等,使一个特征能够从另一个特征接收信息,进一步整合视频片段的运动信息和关键帧的空间信息. 例如,使时序(S)特征传递给空间(T)特征,即由"S"表示"T",计算方式如下:

$$\boldsymbol{Z}_T^0=\boldsymbol{X}_T+e_{\text{pos}} \tag{7}$$

$$\boldsymbol{Z}_{S\to T}^0=\boldsymbol{Z}_T^0 \tag{8}$$

$$\boldsymbol{Z}'^{i}_{S\to T}=\text{CM}^i_{S\to T}\left(\text{LN}\left(\boldsymbol{Z}^{i-1}_{S\to T}\right),\text{LN}\left(\boldsymbol{Z}_S^0\right)\right)+\text{LN}\left(\boldsymbol{Z}^{i-1}_{S\to T}\right) \tag{9}$$

$$\boldsymbol{Z}^i_{S\to T}=f_{\theta^i_{S\to T}}\left(\text{LN}\left(\boldsymbol{Z}'^{i-1}_{S\to T}\right)\right)+\text{LN}\left(\boldsymbol{Z}'^{i-1}_{S\to T}\right) \tag{10}$$

其中,e_{pos}表示一维位置嵌入;f_θ表示由θ参数化的位置前馈子层;LN表示层归一化. 通过特征的潜在自适应表示,特征序列根据低层特征序列信息不断更新,来自不同来源特征的低层信号被转换成一组不同的键值对,以与目标特征交互,从而有利于本文模型保留不同来源特征的信息.

特征提取网络由3个骨干网络组成,即存在3种不同来源的特征(时序特征、空间特征以及光流特征). 因此,时空交叉变压器通过对时序特征(S)、空间特征(T)

以及光流特征(F)两两计算潜在自适应表示,然后分别嵌入Transformer进行交叉表示.

3.3.3 融合注意力模块

将时空交叉变压器的输出进行通道拼接,通过融合注意力关注主要特征抑制次要特征. 为充分聚集特征的时空信息,首先通过两个卷积块进行特征映射,其中,一个1×1卷积层来降低信道维数,一个3×3卷积层来细化语义上下文. 然后,通过一个如图6(a)的上下文注意力模块来根据特征通道间的内部关系产生注意力映射进一步融合特征,最后通过回归与分类进行预测.

3.4 视频动作链接策略

由于已经获得了帧级动作检测,下一步是将这些检测到的边界框链接起来,以在整个视频中构建动作链接. 采用与Kalogeiton等人[27]和Köpüklü等人[44]类似的在线链接算法. 给定一个输入视频流,检测出每一帧的检测结果,假设从连续帧F_t和F_{t+1}中检测到的区域为R_t和R_{t+1},动作c在区域R_t和R_{t+1}中的状态得分为$s_c(R_t)$和$s_c(R_{t+1})$,R_t和R_{t+1}的交集ov并集为重叠,则可以定义动作c的链接得分为

$$s_c(R_t,R_{t+1})=\psi(\mathrm{ov})\cdot\left[s_c(R_t)+s_c(R_{t+1})+\alpha\cdot s_c(R_t)\cdot s_c(R_{t+1})+\beta\cdot\mathrm{ov}(R_t,R_{t+1})\right] \tag{11}$$

其中,$s_c(R_t),s_c(R_{t+1})$为区域R_t和R_{t+1}的类特定分数;ov为这两个区域的并集的交集;α和β为超参数. $\psi(\mathrm{ov})$是一个约束,如果存在重叠(ov>0),则等于1,否则$\psi(\mathrm{ov})$等于0. 使用一个额外的元素$\alpha\cdot s_c(R_t)\cdot s_c(R_{t+1})$扩展了链接分数定义,该元素考虑了两个连续帧之间分数急剧变化的情况,并且能够在实验中提高视频检测的性能. 在计算出所有连接分数后,采用维特比算法寻找生成动作管的最佳路径.

4 实验

在本节中,先对实验数据集、评估指标以及实验条件进行介绍,然后对本文所提方法中各个模块进行消融实验,以分析验证模块的有效性. 最后,将本文方法与最新方法进行比较.

4.1 实验设置

4.1.1 数据集

在两个流行且具有挑战性的时空动作检测数据集UCF101-24[45]和JHMDB-21[46]上,对本文提出的方法进行评估实验.

UCF101-24数据集是UCF101数据集的子集,该数据集由3 207个未剪辑的视频组成,共包含24个运动类别,并带有帧级时空注释. 视频帧的原始分辨率为320×240. 由于频繁的相机抖动、动作实例的移动以及动作持续时间的巨大差异,使得视频级动作检测更具挑战性.

JHMDB-21是HMDB-51数据集的子集. 该数据集由928个视频组成,共包含21个动作类别,同样带有帧级时空注释,视频帧的原始分辨率为320×240. 该数据集面临的挑战包括遮挡、背景混乱和高类间相似性.

两个数据集都提供了3种不同的训练、测试集划分方式,对比以往的工作,按照第一种划分方式进行实验,70%的视频用于训练,30%用于测试,具体情况如表1所示,并使用修正后的注释进行模型训练和评估.

表1 数据集概况

数据集	类别	分辨率	训练集	测试集
UCF101-24	24	320×240	2 275	932
JHMDB-21	21	320×240	660	268

4.1.2 评估指标

采用时空动作检测领域中最流行的两种度量标准,帧级度量平均精度(Frame-mAP)和视频级度量平均精度(Video-mAP)分数来评估动作检测性能. Frame-mAP按照PASCAL VOC 2012度量标准所应用的规则,测量每帧检测的Precision-Recall曲线下的面积.

Precision为准确率,表示模型预测的所有目标中,预测正确检测框数TP占所有预测框数alldetctions的比例. 计算方式如下:

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}=\frac{\mathrm{TP}}{\mathrm{alldetections}} \tag{12}$$

Recall为召回率,表示所有真实目标中,预测正确检测框数TP占所有人工标注框数allgroundtruths的比例,计算公式如下:

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}=\frac{\mathrm{TP}}{\mathrm{allgroundtruths}} \tag{13}$$

Video-mAP度量动作管道生成的精度,当视频帧与实际时空注释的平均IoU大于某个阈值,并且同时正确预测了动作标签,则该检测到的动作管道被视为正确的实例. 最后,计算每个类的平均精度,以及所有类的平均精度.

4.1.3 实验条件

本文所有网络均在PyTorch框架内实现,并且以一种端到端的方式在一张TeslaV100上进行训练. 本文的基线网络架构以ResNext101作为3D骨干网络,Darknet作为2D骨干网络进行训练. 并尝试了分别以ResNet50、ResNext101和R(2+1)D-18为3D骨干网络,Darknet、CSPNet为2D骨干网络的组合作为特征提取网络的检测效果. 在训练过程中,我们使用SmoothL1 Loss、MSELoss、FocalLoss分别计算边界框位置、置信度以及分类损失,并通过Adam优化算法优化网络参数,

初始学习率设为 1×10^{-4},batch size 设置为 12. 对于仅使用 RGB 模态的 SIPD 网络,我们将数据集 RGB 帧的输入序列长度设置为 8、16,大小设置为 224×224,最大 epoch 设置为 15. 对于增加光流模态的 SIPD 网络,本文使用 RAFT[47] 网络提取光流信息,最大 epoch 设置为 20. 其他参数均采用 Pytorch 框架的默认值. 本文的模型在 UCF101-24 数据集上训练时间约为 5~7 天,在 JHMDB-21 数据集上的训练时间约为 1 天.

4.2　实验结果及分析

在本节中,首先探索不同的骨干网络作为特征提取网络对动作检测的影响,并通过不同的特征提取网络验证本文方法的有效性. 最后,以 ResNext101 作为 3D 骨干网络 Darknet 作为 2D 骨干网络在 UCF101-24、JHMDB-21 两个数据集上进行消融实验.

4.2.1　骨干网络的选择

以往的研究[44]表明,单独 2D 和 3D CNN 并不能很好地解决动作检测任务,2D-CNN 学习更精细的空间特征,3D-CNN 更专注于运动过程,两者结合才能够更好地对时空信息进行建模. 因此,本文通过尝试不同 2D、3D 骨干网络的组合作为特征提取网络来提升动作检测的精度. 如表 2 所示,利用 Frame-mAP 来评估模型的精度. 通过比较发现,R(2+1)D 网络作为 3D 骨干网络,CSPNet 网络作为 2D 骨干网络的组合,能够有效地对时空信息进行建模,提高动作检测的准确率,较基准模型在 UCF101-24 数据集上平均提升了 1.22%,在 JHMDB-21 数据集上平均提升了 4.81%.

表 2　骨干网络的选择效果的比较　　单位:%

骨干网络	输入帧数	数据集	
		UCF101-24	JHMDB-21
ResNext101+Darknet	8	79.40	64.61
ResNext101+Darknet	16	80.93	72.31
ResNext101+CSPNet	8	79.76(+0.36)	67.50(+2.89)
ResNext101+CSPNet	16	81.24(+0.31)	72.61(+0.30)
R(2+1)D+Darknet	8	80.20 (+0.80)	64.84(+0.23)
R(2+1)D+Darknet	16	81.26(+0.33)	75.78(+3.47)
R(2+1)D+CSPNet	8	80.59(+1.16)	68.24(+3.63)
R(2+1)D+CSPNet	16	81.72(+0.79)	76.21(+3.90)

4.2.2　乱序重排划分片段选取及性能分析

本文提出的乱序重排模块根据输入视频长度,将其划分为 1~16 个视频片段. 当片段数量等于输入视频长度时,片段内视频帧数量过少,每个片段仅包含静态信息,而不包含时间信息,难以对时间序列进行建模.

为了研究划分片段数量的选取对实验结果的影响,以 ResNext101+Darknet 的组合作为骨干网络,分别对输入长度为 8、16 的视频选取 1、2、4、8 作为片段数量进行划分,并在 JHMDB-21 数据集上进行实验对比. 实验结果如表 3 所示,加粗数据表示最优结果,可以看出,当划分片段数量为 2 时,检测准确率最高,划分片段长度能够很好地表示时序信息. 随着帧集数量增大,帧集包含的时间信息逐渐减少,准确率逐渐降低. 因此,本文在实验中为对输入视频划分为 2 个子片段进行短期时间建模.

表 3　选取不同划分片段数量结果的比较

骨干网络	输入帧数	划分片段数	帧级 mAP/%
ResNext101+Darknet	8	1	64.61
ResNext101+Darknet	8	2	**67.88**
ResNext101+Darknet	8	4	66.75
ResNext101+Darknet	8	8	66.59
ResNext101+Darknet	16	1	72.31
ResNext101+Darknet	16	2	**73.75**
ResNext101+Darknet	16	4	73.43
ResNext101+Darknet	16	8	73.07

此外,以 ResNet50+Darknet、ResNext101+Darknet、R(2+1)D+CSPNet 这 3 个组合作为特征提取网络,在此基础上增加了乱序重排模块分别进行实验,进一步分析本文提出的乱序重排模块的贡献. 实验结果如表 4 所示,可以看出乱序重排模块的加入在所有主干网和数据集上的性能都优于基准模型,验证了方法的有效性.

表 4　乱序重排模块效果比较　　单位:%

骨干网络	乱序重排	数据集	
		UCF101-24	JHMDB-21
ResNet50+Darknet	×	77.49	60.23
	✓	80.33(+2.70)	62.52(+2.29)
ResNext101+Darknet	×	80.93	72.31
	✓	81.05(+0.12)	73.75(+1.44)
R(2+1)D+CSPNet	×	81.72	76.21
	✓	82.05(+0.33)	77.70(+1.59)

4.2.3　关键帧光流性能分析

图 5 显示了基于动作表示的关键帧光流与全光流的运动特征的表示,可以看出对所有相邻帧提取的光流存在着噪声数据,通过运动表示关键帧光流信息,能够使网络更好地关注运动区域. 为了进一步分析基于动作表示的关键帧光流的贡献,本文在基准模型的基础上增加了光流分支,并对关键帧光流与全光流进行比较. 分别以 ResNet50+Darknet+Darkne, ResNext101+Darknet+Darknet, R(2+1)D+CSPNet+CSPNet 作为特征提取网络在 JHMDB-21 数据集上进行实验,实验结果如表 5 所示,加粗数据表示最优结果. 可以看出,与全视频提取光流仅对关键帧提取光流信息相比,有效地提高了模型的检测精度.

4.2.4　多尺度时空交叉注意力性能分析

根据本文提出的多尺度时空交叉注意力的组成,

表5 基于动作表示的关键帧光流效果比较

骨干网络	光流使用情况	Frame-mAP/%
ResNet50+Darknet+Darknet	全部光流	62.85
	关键帧光流	63.77
ResNext101+Darknet+Darknet	全部光流	69.39
	关键帧光流	72.89
R(2+1)D+CSPNet+CSPNet	全部光流	75.12
	关键帧光流	79.00

分别在基准模型的基础上增加了多尺度注意力网络、时空交叉变压器以及融合注意力，来分析本文提出的多尺度时空交叉注意力模块的贡献．以ResNet50+Darknet、ResNext101+Darknet、R（2+1）D+CSPNet分别作为特征提取网络进行实验．实验结果如表6所示，可以看出各个模块的加入在所有特征提取网络和数据集上的性能都优于基准模型，验证了本文提出的多尺度时空交叉注意力模块可以发现更具辨别力的时空表示，给网络检测性能带来了显著的提升，并且可以很好地推广到不同的主干网络和数据集．

表6 多尺度时空交叉注意力组成模块效果 单位：%

骨干网络	多尺度注意力网络	数据集	
		UCF101-24	JHMDB-21
ResNet50 + Darknet	×	77.49	60.23
	✓	78.39(+0.90)	62.81(+2.58)
ResNext101 + Darknet	×	80.93	72.31
	✓	82.05(+1.12)	72.58(+0.27)
R(2+1)D + CSPNet	×	81.72	76.21
	✓	82.14(+0.42)	76.35(+0.14)
骨干网络	时空交叉变压器	UCF101-24	JHMDB-21
ResNet50 + Darknet	×	77.49	60.23
	✓	78.38(+0.89)	61.85(+1.62)
ResNext101 + Darknet	×	80.93	72.31
	✓	80.96(+0.03)	75.14(+2.83)
R(2+1)D + CSPNet	×	81.72	76.21
	✓	82.07(+0.35)	77.87(+1.66)
骨干网络	融合注意力	UCF101-24	JHMDB-21
ResNet50 + Darknet	×	77.49	60.23
	✓	78.79(+1.30)	63.24(+2.99)
ResNext101 + Darknet	×	80.93	72.31
	✓	81.82(+0.68)	74.18(+1.87)
R(2+1)D + CSPNet	×	81.72	76.21
	✓	82.59(+0.87)	78.83(+2.62)

4.2.5 消融实验

以ResNext101作为3D骨干网络Darknet作为2D骨干网络在UCF101-24以及JHMDB-21上进行了消融实验．表7总结了所提方法的每个组成部分的贡献，从这些结果中可以看出，每个组成部分确实相较基准模型带来了性能增益．

最后，对多尺度时空交叉注意力的模型复杂度进行分析，并以消融实验的形式展示．表8总结了多尺度时空交叉注意力各个模块的计算量和参数量．可以看出，每个模块的增加带来的计算量和参数量的增幅相对较小，本文的改进以较小的计算量和参数量增幅就能够大幅度提升检测精度．

4.2.6 收敛性分析

图7展示了本文的方法在2个数据集上各个损失函数值与iterations的关系图．其中，横轴表示iteration，纵轴表示损失．从结果可以发现，在整个训练过程中，目标函数的值单调递减，并且在两个数据集上收敛．目标函数值在10个epoch后趋于稳定，说明采用随机梯度下降优化算法Adam可以有效地训练．

4.3 与近几年的主流方法的比较

在本节中，以R(2+1)D网络作为3D骨干网络，CSPNet网络作为2D骨干网络的组合将所提出的方法（SIPD）与当前动作检测任务中最先进的方法进行比较．

表7　模块消融结果　　单位:%

乱序重排	多尺度时空交叉注意力			输入帧数	数据集	
	融合注意力模块	多尺度注意力网络	时空交叉变压器		UCF101-24	JHMDB-21
				8	79.40	64.61
				16	80.93	72.31
✓				8	80.73(+1.33)	67.88(+3.21)
				16	81.05(+0.12)	73.75(+1.44)
✓	✓			8	80.26(+0.86)	69.07(+4.46)
				16	81.32(+0.40)	74.76(+2.45)
✓		✓		8	80.58(+1.18)	68.39(+3.78)
				16	81.94(+1.01)	72.93(+0.62)
✓		✓	✓	8	80.04(+0.64)	68.43(+3.82)
				16	81.64(+0.71)	74.42(+2.11)
✓	✓	✓	✓	8	80.99(+1.59)	69.94(+5.33)
				16	82.10(+1.17)	74.59(+2.28)

表8　模型复杂度分析

多尺度时空交叉注意力			计算量/G	参数量/M
多尺度注意力网络	时空交叉变压器	融合注意力模块		
✓			0.75	23.39
✓	✓		3.31	75.35
✓	✓	✓	4.40	113.11

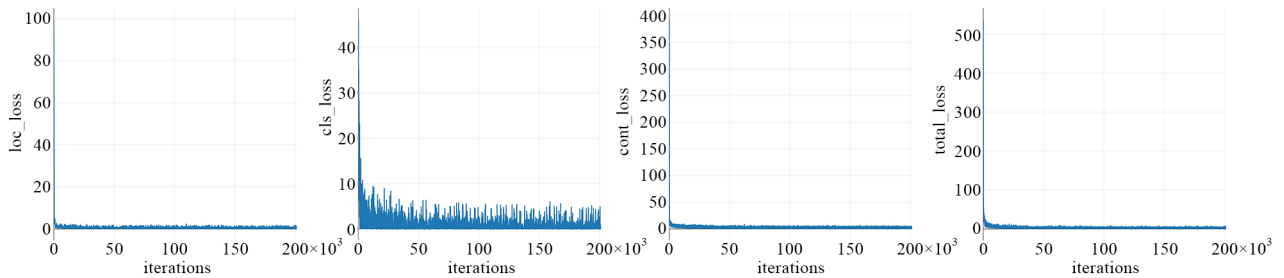
4.3.1　帧级实验对比

首先在UCF101-24数据集上,根据Frame-mAP动作检测精度与最先进的方法进行比较. 表9为UCF101-24数据集上IoU=0.5时各个方法Frame-mAP检测精度的比较,加粗数据表示最优结果. 值得注意的是,本文方法仅使用RGB输入就能优于近几年主流的方法.

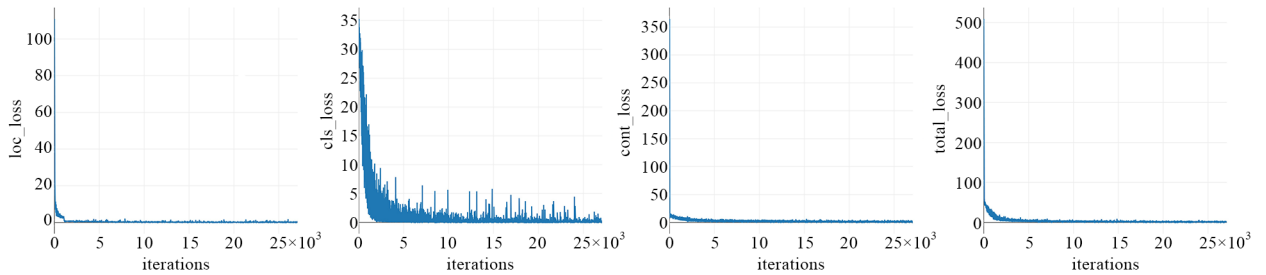
表10为JHMDB-21数据集上IoU为0.5时各个方法Frame-mAP检测精度的比较,加粗数据表示最优结果. 可以看出,本文仅基于RGB的输入就能表现出优于现有双流特征的方法,针对JHMDB-21数据集存在高类间相似性易于混淆的难样本数据. 本文还融合了基于动作表示的关键帧光流,进一步提高了准确率,并表现出优于近几年主流方法的检测性能.

4.3.2　视频级实验对比

此外,本文还评估了SIPD仅使用RGB输入的Video-mAP以及在单个GPU上的模型运行速度. 其中,模型运行速度是基于每帧处理时间来评估的,通过取每个视频的运行时间并除以输入长度t得到. 在图8中,比较了在JHMDB-21数据集上的现有包含运行速度



(*a*) UCF101-24数据集



(*b*) J-HMDB-21数据集

图7　模型收敛性分析

表9 UCF101-24数据集帧级mAP与近几年主流方法的比较

方案	是否使用光流	mAP/%
T-CNN[23]	否	41.40
ACDnet[24]	否	70.92
ACT[27]	否	69.50
TEDdet[31]	否	64.70
STMA[48]	否	78.80
YOWO*[44]	否	80.40
ACAR-NET[6]	否	84.30
Zhang et al.[49]	是	67.70
Pramono[50]	是	73.70
STEP[30]	是	75.00
MOC[28]	是	78.00
SIPD(本文方法)	否	**84.71**

表10 JHMDB-21数据集帧级mAP与近几年主流方法的比较

方案	是否使用光流	mAP/%
T-CNN[23]	否	61.30
ACDnet[24]	否	49.53
ACT[27]	否	65.70
TEDdet[31]	否	70.80
STMA[48]	否	77.60
YOWO*[44]	否	72.31
SIPD(本文方法)	否	78.40
Zhang et al.[49]	是	37.40
Pramono[50]	是	76.70
MOC[28]	是	70.80
SIPD(本文方法)	是	**79.00**

和检测精度的方法(具体数值来自原论文). 可以看出,本文方法在速度以及精度方面的综合表现优于现有其他方法.

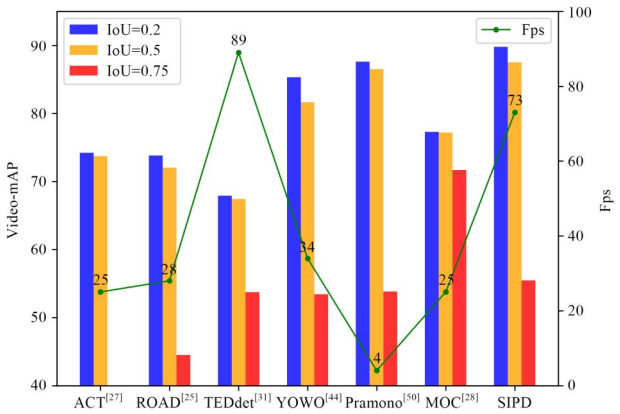


图8 模型Video-mAP与性能的比较

4.4 可视化展示

4.4.1 特征可视化

为了直观地了解本文方法的效果,图9展示了多尺度时空交叉注意力模块的效果. 本文采用Grad-CAM算法[51]来生成类激活图,它突出了网络在识别特定类时的重要区域. 本文将仅使用2D卷积特征、3D卷积特征,2D与3D卷积特征简单拼接以及多尺度时空交叉注意力模块进行比较. 可以看出,单独使用2D、3D卷积并不能很好地关注运动区域,而与通过通道融合2D、3D卷积特征相比,本文网络更好地关注运动动作相关的重要区域.

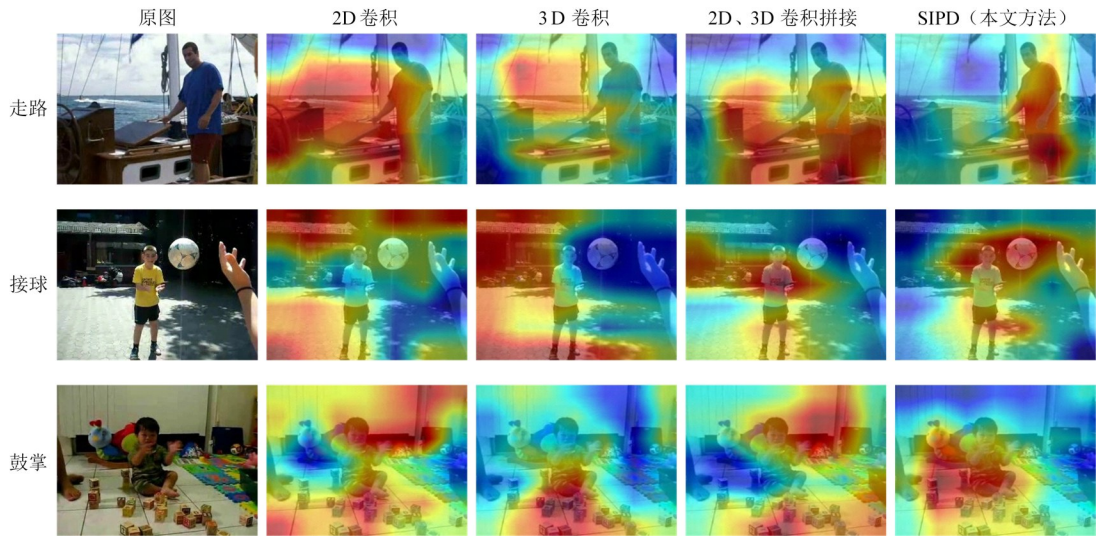


图9 多尺度时空交叉注意力关注效果对比可视化

4.4.2 模型效果可视化

图10展示了本文方法在低光照、小目标、局部遮挡等困难场景下的检测效果. 可以看出,在这些困难场景下仅根据当前帧很难识别出动作的类型,如第1行所示走路、跑步与起立3个动作,仅能艰难地看出一个站立的身体影,无法判断其运动类别. 再如第3行所示梳

头、射击、爬楼梯3个动作,只能看到身体的部分区域,并不能看到整体的运动.针对这种情况,本文方法会根据输入前后帧内容来进行正确的动作检测.而如第2行的接球、跳远、高尔夫3个动作相对其他动作发生场景范围大、运动目标小,针对这种情况,本文方法通过提取多尺度目标特征有效地检测出此类小目标动作.总的来说,本文方法在多种困难场景下都展现出一定的稳定性.

图10　帧级检测示例图

图11展示了本文方法根据帧级检测结果生成动作链接的效果.从图11中高尔夫、跳远、引体向上等动作的链接效果可以看出,本文的链接算法能够有效地利用帧级检测结果,根据动作类别以及运动区域来进行正确的动作定位.

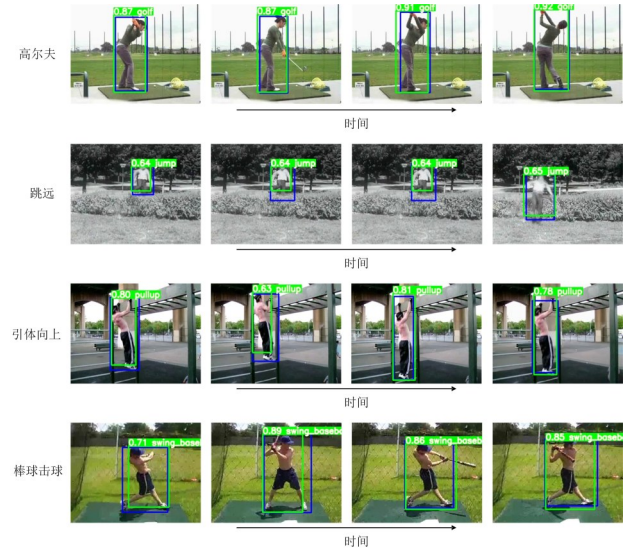


图11　视频检测示例图

此外,结合时事在网上搜集了一些实际生活中的视频,如世界杯、冬奥会、NBA以及一些综艺中的视频进行测试,图12展示了本文方法在这些场景中的检测效果.可以看出,本文方法不仅能够准确地检测出真实的运动项目,如图中世界杯踢足球、冬奥会滑雪、NBA篮球扣篮、奥运会跳水等动作,且对于如跑步、梳头等日常生活的动作也能够进行准确的检测.

图12　真实场景测试示例图

5　结论

时序信息与空间信息在时空动作检测中都起到十分重要的作用,如何有效地感知融合时序信息与空间信息构建运动表示是时空动作检测的一个挑战.本文重点对基于时空特征表示的实时动作检测进行了研究,提出了一种基于时空交叉感知的实时动作检测方法.简单来说,针对单一尺度时空特征描述性不足,提出一个多尺度注意力网络来学习长期的时间依赖和空间上下文信息.针对时序和空间两种不同来源特征的融合,提出了一种新的运动显著性增强融合策略.最后,针对JHMDB-21数据集存在高类间相似性与难样本

数据易于混淆等问题,提出了基于动作表示的关键帧光流动作检测方法,并且通过公共数据集验证了本文提出的动作检测方法的有效性. 然而,现有的动作检测模型大多基于2D与3D骨干网络的结合,而3D骨干网络所带来的计算量也相对较大. 如何减少3D骨干网络计算量、寻求更轻量化的方法可作为下一步的研究方向.

参考文献

[1] SHAO D, ZHAO Y, DAI B, et al. FineGym: A hierarchical video dataset for fine-grained action understanding[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2613-2622.

[2] 罗会兰, 童康, 孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报, 2019, 47(5): 1162-1173.
LUO H L, TONG K, KONG F S. The progress of human action recognition in videos based on deep learning: A review[J]. Acta Electronica Sinica, 2019, 47(5): 1162-1173. (in Chinese)

[3] 杨珂, 王敬宇, 戚琦, 等. LSCN: 一种用于动作识别的长短时序关注网络[J]. 电子学报, 2020, 48(3): 503-509.
YANG K, WANG J Y, QI Q, et al. LSCN: Concerning long and short sequence together for action recognition[J]. Acta Electronica Sinica, 2020, 48(3): 503-509. (in Chinese)

[4] XU M Z, XIONG Y J, CHEN H, et al. Long short-term transformer for online action detection[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2021: 1086-1099.

[5] DAI R, DAS S, KAHATAPITIYA K, et al. MS-TCT: Multi-scale temporal ConvTransformer for action detection [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 20009-20019.

[6] PAN J T, CHEN S Y, SHOU M Z, et al. Actor-context-actor relation network for spatio-temporal action localization [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 464-474.

[7] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 1933-1941.

[8] ZHAO J J, SNOEK C G M. Dance with flow: Two-In-one stream action detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 9927-9936.

[9] LI H H, JIANG X D, GUAN B L, et al. Joint feature optimization and fusion for compressed action recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 7926-7937.

[10] SHOU Z, LIN X D, KALANTIDIS Y, et al. DMC-net: Generating discriminative motion cues for fast compressed video action recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 1268-1277.

[11] TAO L, WANG X T, YAMASAKI T. Rethinking motion representation: Residual frames with 3D ConvNets[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2021, 30: 9231-9244.

[12] 桑海峰, 赵子裕, 何大阔. 基于循环区域关注和视频帧关注的视频行为识别网络设计[J]. 电子学报, 2020, 48(6): 1052-1061.
SANG H F, ZHAO Z Y, HE D K. Recurrent region attention and video frame attention based video action recognition network design[J]. Acta Electronica Sinica, 2020, 48(6): 1052-1061. (in Chinese)

[13] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5534-5542.

[14] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6450-6459.

[15] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 2625-2634.

[16] LI Y H, SONG S J, LI Y Q, et al. Temporal bilinear networks for video action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8674-8681.

[17] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 7082-7092.

[18] LI Y, JI B, SHI X T, et al. TEA: Temporal excitation and aggregation for action recognition[C]//2020 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 906-915.

[19] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. New York: ACM, 2014: 568-576.

[20] LIU X L, WANG Q M, HU Y, et al. End-to-end temporal action detection with transformer[J]. IEEE Transactions on Image Processing, 2022, 31: 5427-5441.

[21] MIRIAM JACOB G, STENGER B. Facial action unit detection with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 7676-7685.

[22] GKIOXARI G, MALIK J. Finding action tubes[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 759-768.

[23] HOU R, CHEN C, SHAH M. Tube convolutional neural network (T-CNN) for action detection in videos[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5823-5832.

[24] LIU Y, YANG F, GINHAC D. ACDnet: An action detection network for real-time edge computing based on flow-guided feature approximation and memory aggregation [J]. Pattern Recognition Letters, 2021, 145: 118-126.

[25] SINGH G, SAHA S, SAPIENZA M, et al. Online real-time multiple spatiotemporal action localisation and prediction[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 3657-3666.

[26] SAHA S M, SINGH G, SAPIENZA M, et al. Deep learning for detecting multiple space-time action tubes in videos[C]//Proceedings of the British Machine Vision Conference. New York: British Machine Vision Association, 2016: 58.

[27] KALOGEITON V, WEINZAEPFEL P, FERRARI V, et al. Action tubelet detector for spatio-temporal action localization[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4415-4423.

[28] LI Y X, WANG Z X, WANG L M, et al. Actions as moving points[C]//European Conference on Computer Vision. Cham: Springer, 2020: 68-84.

[29] KUMAR A, RAWAT Y S. End-to-end semi-supervised learning for video action detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14680-14690.

[30] YANG X T, YANG X D, LIU M Y, et al. STEP: Spatio-temporal progressive learning for video action detection [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 264-272.

[31] LIU Y, YANG F, GINHAC D. TEDdet: Temporal feature exchange and difference network for online real-time action detection[J]. IEEE Access, 2022, 10: 37870-37881.

[32] 胡正平, 刁鹏成, 张瑞雪, 等. 3D多支路聚合轻量网络视频行为识别算法研究[J]. 电子学报, 2020, 48(7): 1261-1268.
HU Z P, DIAO P C, ZHANG R X, et al. Research on 3D multi-branch aggregated lightweight network video action recognition algorithm[J]. Acta Electronica Sinica, 2020, 48(7): 1261-1268. (in Chinese)

[33] WANG Z W, SHE Q, SMOLIC A. ACTION-net: Multipath excitation for action recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 13209-13218.

[34] PRAMONO R R A, CHEN Y T, FANG W H. Spatial-temporal action localization with hierarchical self-attention[J]. IEEE Transactions on Multimedia, 2021, 24: 625-639.

[35] 罗会兰, 王婵娟. 行为识别中一种基于融合特征的改进VLAD编码方法[J]. 电子学报, 2019, 47(1): 49-58.
LUO H L, WANG C J. An improved VLAD coding method based on fusion feature in action recognition[J]. Acta Electronica Sinica, 2019, 47(1): 49-58. (in Chinese)

[36] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.

[37] YUE K Y, SUN M, YUAN Y C, et al. Compact generalized non-local network[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 6511-6520.

[38] CAO Y, XU J R, LIN S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 1971-1980.

[39] CHEN Y P, KALANTIDIS Y, LI J S, et al. A2-nets: Double attention networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018:

350-359.
[40] LI X, ZHONG Z S, WU J L, et al. Expectation-maximization attention networks for semantic segmentation[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9166-9175.
[41] CHEN W L, ZHU X G, SUN R Q, et al. Tensor low-rank reconstruction for semantic segmentation[C]//European Conference on Computer Vision. Cham: Springer, 2020: 52-69.
[42] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.
[43] CHEN Q, WANG Y M, YANG T, et al. You only look one-level feature[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 13034-13043.
[44] KÖPÜKLÜ O, WEI X Y, RIGOLL G. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization[EB/OL]. (2019-11-15)[2022-07-16]. https://arxiv.org/abs/1911.06644v5.
[45] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[EB/OL]. (2012-12-03)[2022-07-16]. https://arxiv.org/abs/1212.0402.
[46] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition[C]// 2011 International Conference on Computer Vision. Piscataway: IEEE, 2011: 2556-2563.
[47] TEED Z, DENG J. RAFT: Recurrent all-pairs field transforms for optical flow[C]//European Conference on Computer Vision. Cham: Springer, 2020: 402-419.
[48] ZHANG H C, ZHAO X. Spatio-temporal motion aggregation network for video action detection[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 2180-2184.
[49] ZHANG D J, HE L C, TU Z G, et al. Learning motion representation for real-time spatio-temporal action localization[J]. Pattern Recognition, 2020, 103: 107312.
[50] PRAMONO R R A, CHEN Y T, FANG W H. Hierarchical self-attention network for action localization in videos [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 61-70.
[51] DIBA A, SHARMA V, VAN GOOL L. Deep temporal linear encoding networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 1541-1550.

作者简介

柯 逍 男,1983年生,福建福州人. 博士,福州大学教授、博士生导师. 主要研究方向为计算机视觉、模式识别.
E-mail: kex@fzu.edu.cn

缪 欣 女,1997年生,福建福安人. 福州大学计算机与大数据学院硕士研究生. 主要研究方向为计算机视觉、动作识别.
E-mail: 200320077@fzu.edu.cn

郭文忠 男,1979年生,福建惠安人. 博士,福州大学教授,博士生导师. 主要研究方向为计算智能及其应用.
E-mail: guowenzhong@fzu.edu.cn