中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)05-1499-14

论文引用格式:He C J, Liu Q Y and Wang Z L. 2023. Modeling interaction and profiling dependency-relevant video action detection. Journal of Image and Graphics, 28(05):1499-1512(贺楚景, 刘钦颖, 王子磊. 2023. 建模交互关系和类别依赖的视频动作检测. 中国图象图形学报, 28(05):1499-1512)[DOI:10.11834/jig.211040]

建模交互关系和类别依赖的视频动作检测

贺楚景[1],刘钦颖[2],王子磊[1,2*]

1. 中国科学技术大学大数据学院,合肥　230027; 2. 中国科学技术大学自动化系,合肥　230027

摘　要:**目的**　视频动作检测旨在检测出视频中所有人员的空间位置,并确定其对应的动作类别。现实场景中的视频动作检测主要面临两大问题,一是不同动作执行者之间可能存在交互作用,仅根据本身的区域特征进行动作识别是不准确的;二是一个动作执行者在同一时刻可能有多个动作标签,单独预测每个动作类忽视了它们的内在关联。为此,本文提出了一种建模交互关系和类别依赖的视频动作检测方法。**方法**　首先,特征提取部分提取出关键帧中每个动作执行者的区域特征;然后,长短期交互部分设计短期交互模块(short-term interaction module,STIM)和长期交互模块(long-term interaction module,LTIM),分别建模动作执行者之间的短期时空交互和长期时序依赖,特别地,基于空间维度和时间维度的异质性,STIM采用解耦机制针对性地处理空间交互和短期时间交互;最后,为了解决多标签问题,分类器部分设计类别关系模块(class relationship module,CRM)计算类别之间的依赖关系以增强表征,并利用不同模块对分数预测结果的互补性,提出一种双阶段分数融合(two-stage score fusion,TSSF)策略更新最终的概率得分。**结果**　在公开数据集AVA v2.1(atomic visual actions version 2.1)上进行定量和定性分析。定量分析中,以阈值取0.5时的平均精度均值mAP@IoU 0.5(mean average precision @ intersection over union 0.5)作为主要评价指标,本文方法在所有测试子类、Human Pose大类、Human-Object Interaction大类和Human-Human Interaction大类上的结果分别为31.0%,50.8%,22.3%,32.5%。与基准模型相比,分别提高了2.8%,2.0%,2.6%,3.6%。与其他主流算法相比,在所有子类上的指标相较于次好算法提高了0.8%;定性分析中,可视化结果一方面表明了本文模型能精准捕捉动作执行者之间的交互关系,另一方面体现了本文在类别依赖建模上的合理性和可靠性。此外,消融实验证明了各个模块的有效性。**结论**　本文提出的建模交互关系和类别依赖的视频动作检测方法能够提升交互类动作的识别效果,并在一定程度上解决了多标签分类问题。

关键词:视频动作检测;多标签分类;交互关系建模;双阶段融合;深度学习;注意力机制

Modeling interaction and profiling dependency-relevant video action detection

He Chujing[1], Liu Qinying[2], Wang Zilei[1,2*]

1. School of Big Data, University of Science and Technology of China, Hefei　230027, China;

2. Department of Automation, University of Science and Technology of China, Hefei　230027, China

Abstract: **Objective**　Video action detection is one of the challenging tasks in computer vision and video recognition, which aims to locate all actors and recognize their actions in video clips. However, two major problems are required to be

收稿日期:2021-11-04;**修回日期**:2022-03-01;**预印本日期**:2022-03-08

*通信作者:王子磊　zlwang@ustc.edu.cn

基金项目:国家自然科学基金项目(62176246,61836008)

Supported by:National Natural Science Foundation of China(62176246,61836008)

resolved in the real world: First, there are pairwise interactions between actors in real scenes, so it is suboptimal to perform action recognition only in terms of their own regional features. Specifically, explicitly modeling interactions may be beneficial to the performance of action detection. Second, action detection is a multi-label classification task because an actor may perform multiple types of actions at the same time. We argue that it is beneficial to consider the inherent dependency between different classes. In this study, we propose a video action detection framework that simultaneously modeling interaction between actors and dependency between categories. **Method** Our framework proposed consists of three main parts: actor feature extraction, long short-term interaction and classification. In detail, the actor feature extraction part first utilizes the Faster region based convolutional neural network (Faster R-CNN) as a person detector to detect the potential actors in the whole dataset. After this, we only keep the detected actors which have relatively high confidence score. Furthermore, we take the SlowFast network as backbone to extract features from the raw videos. For each actor, we apply the RoIAlign operation on the extracted feature map based on the location of the actor and the corresponding feature of the actor is generated. To include the geometric information, the coordinates of actors is embedded into their features. The actor-related features are as the input of following steps. The long short-term interaction part includes short-term interaction module (STIM) and long-term interaction module (LTIM). The STIM can leverage the graph attention network (GAT) to model short-term spatio-temporal interactions between actors, and LTIM can use long-term feature banks (LFB) to model long-term temporal dependency between actors. For short-term spatio-temporal interaction modeling, the intersection over union (IoU)-tracker is used to connect the bounding boxes of the same person in temporal. For optimization, we propose a decoupling mechanism to handle spatial and temporal interactions. The interaction between nodes of different actors at the same time step is defined as spatial interaction, and the interaction between nodes of the same actor at different time steps is recognized as temporal interaction. The graph attention network (GAT) is applied for each of them, where the nodes are the actor features, and their pairwise relationships are then represented via the edges. For long-term temporal dependency modeling, a sliding window mechanism is used to obtain a long-term feature bank, which can contain a large scale of temporal contexts. Current short-term features and the long-term feature bank can be transmitted into the non-local module, where the short-term features are used as queries, and the features in the long-term feature bank are used as key-value pairs to extract relevant long-term temporal context information. The classification part is based on a class relationship module (CRM), which first extracts class-specific feature for each class of each actor. The features of different classes of the same actor can be passed into the self-attention module to compute the semantic correlation between action classes. Finally, we propose a two-stage score fusion (TSSF) strategy to update the classification probability scores on the basis of the complementarity of different modules. **Result** We carry out quantitative and qualitative analysis on the public dataset atomic visual actions version 2.1 (AVA v2.1). For the quantitative analysis, the evaluation metrics is the average precision (AP) with an IoU threshold of 0.5. For each class, we compute the average precision and report the average over all classes. The results of our method in all test sub-categories, Human Pose category, Human-Object Interaction category, and Human-Human Interaction category are 31.0%, 50.8%, 22.3%, and 32.5%, respectively. Compared to the baseline, it increases by 2.8%, 2.0%, 2.6% and 3.6% each. Compared to actor-centric relation network (ACRN), video action transformer network (VAT), actor-context-actor relation network (ACAR-Net) and other related algorithms, our method is optimized by 0.8% further. For the qualitative analysis, the visualization results demonstrate that our method can capture the interaction accurately between action performers, and the rationality and reliability of the class dependency modeling can be reflected as well. To verify the effectiveness of the proposed modules, a series of ablation experiments are conducted. Additionally, the methods proposed are all end-to-end training, but the method proposed in this paper uses a fixed backbone, which has a faster training speed and lower computing resource consumption. **Conclusion** To fully interpret the interaction between actors and dependency between classes, a video action detection framework is illustrated. The effectiveness of our proposed method is validated according to the experimental results on AVA v2.1 dataset.

Key words: video action detection; multi-label classification; interaction relationship modeling; two-stage fusion; deep learning; attention mechanism

0 引　言

随着电子拍摄设备的普及，产生了海量的视频数据，绝大部分包含以人为主体的动作。这些视频已经广泛应用于智能机器人、安防监控和自动驾驶等多个领域(罗会兰 等，2019)。如何对海量视频中的内容进行理解和分析成为热点课题。视频动作检测是其中的一项关键技术，是指给定一段视频，需要检测出视频中所有人员的空间位置，并确定其对应的动作类别，在现实场景下具有极大的应用价值与研究意义。

针对视频动作检测，大部分研究工作分两步进行。首先对动作执行者进行帧级的目标检测，获取关键帧中一系列检测框的位置坐标和动作性分数；然后使用I3D(inflated 3d convnet)(Carreira和Zisserman，2017)等3维卷积神经网络处理以关键帧为中心的视频片段，提取全局的时空表征，再基于检测框对应的区域特征进行动作分类。上述流程中，针对图像的目标检测已经非常成熟，可以达到理想的性能，因此该任务的主要难点在于动作识别部分。时序信息对于动作的识别起着至关重要的作用，为了充分地利用时序信息以及更好地提取视频表征，SlowFast(Feichtenhofer等，2019)是一种新型的骨干网络，对视频的空间维度和时间维度分别进行处理，即以两种不同的帧率运行的单流体系结构。其中，具有低帧率、低时间分辨率的Slow pathway用于捕捉空间语义，而高帧率、高时间分辨率的Fast pathway能够以精细的时间分辨率捕捉运动信息，该网络显著提高了一般场景下视频动作检测的性能。

然而，在实际应用中，由于很多动作类别具有交互性质，即参与动作的主体不只有动作执行者，还包括周围环境上下文，所以仅根据动作执行者本身的区域特征进行动作推断是不准确的。最近很多方法致力于对动作执行者(actor)与其周围上下文(context)之间的交互作用进行建模，例如场景、其他人和物体等。ACRN(actor-centric relation network)(Sun等，2018)是以动作执行者为中心的关系网络，从全局场景中生成成对关系特征。VAT(video action transformer network)(Girdhar等，2019)将Transformer网络(Vaswani等，2017)引入到视频动作检测任务上，通过注意力机制(attention mechanism)关注与动作执行者相关的周围时空上下文。为了利用时序上的全局信息，LFB(long-term feature banks)(Wu等，2019)建立长期特征库为模型提供长期时间支持，以计算动作执行者之间的远程交互。AIA(asynchronous interaction aggregation network)(Tang等，2020)构造了一种异步交互聚合网络，试图聚合多个类型的交互作用，利用不同类型交互块之间的深度嵌套来增强目标特征。ACAR-Net(actor-context-actor relation network)(Pan等，2021)提出捕捉间接的高阶支持信息，从而有效地推理人物在复杂场景中的行为。这些方法的基本研究思路都是围绕动作执行者对潜在的交互关系进行建模，没有针对性地处理空间交互和时间交互，忽略了空间维度和时间维度的异质性。具体来说，空间交互往往体现在肢体动作的互动上，强调交互对象之间的空间相对位置关系；而时间交互注重的是短期时间内同一个动作的连续性，以及长期时间内不同动作的关联性。因此，本文对动作执行者之间的空间交互、短期时间交互和长期时间交互分别建模。

此外，在现实场景中，一个动作执行者可能同时执行多种类别的动作。针对多标签分类问题，现有模型通常使用二元交叉熵(binary cross entropy loss，BCELoss)组合作为损失函数，操作简单且效果较好，获得了广泛应用。但是这种实现方式意味着直接将多标签分类问题视作多个类上的二分类问题，最终预测时各个类别是独立进行的，没有考虑动作类之间的内在关联。对本文的数据集AVA v2.1(atomic visual actions version 2.1)(Gu等，2018)而言，有些动作对的共现概率非常高，例如hug和kiss、stand和watch a person等，说明对应的动作类之间可能存在较强的语义相关性。反之，有些动作对的共现概率非常低，例如hand shake和kick、swim和drink等，则对应动作类别之间的相关性也较低。基于这种现象，本文认为考虑不同动作类之间的依赖关系可以对多标签分类起到一定的辅助作用。

为了解决上述问题，本文提出了一种同时考虑交互关系和类别依赖的视频动作检测方法，对动作执行者之间的交互作用以及动作类别之间的语义相关性进行建模。本文方法的大致流程如下：首先提取动作执行者对应的区域特征；然后短期交互模块使用两层图注意力网络(graph attention network，GAT)(Velickovic等，2018)分别对空间交互和短期时间交互进行建模，获取局部上下文信息；同时在长

期交互模块中引入长期特征库(long-term feature banks, LFB)(Wu等, 2019),与短期特征进行融合,提取与之相关的全局上下文信息;最后设计类别关系模块(class relationship module, CRM),计算不同动作类之间的语义相关性对原始的类别特征进行加权,得到增强后的各个类表征,并提出一种双阶段分数融合(two-stage score fusion, TSSF)策略更新最终的概率得分。本文的主要贡献如下:1)针对性地对动作执行者之间的空间交互、短期时间交互和长期时间交互进行建模,以增强目标特征的表达能力,提升交互类动作的识别效果,既考虑了空间维度和时间维度的异质性,又兼顾了时序上的局部信息和全局信息;2)对于多标签问题,设计类别关系模块挖掘不同动作类之间的依赖关系,并利用这种关系对原始类别表征进行融合,使得学习到的表征具有更强的鲁棒性和区分度,进一步提高了多标签分类的准确性;3)在数据集AVA v2.1上实验评估了本文方法,定量和定性分析说明了本文方法的先进性,消融实验验证了各个模块的有效性。

1 相关方法

1.1 视频中的关系推理

关系推理在视频理解领域扮演着十分重要的角色,因为在一些复杂场景中,要想精确地识别一个动作执行者的行为往往需要考虑其与其他对象之间的关系。Sun等人(2018)提出以动作执行者为中心的关系网络ACRN(actor-centric relation network),模型从动作执行者和全局场景特征中聚合成对关系信息,生成用于动作分类的关系特征。Wu等人(2019)提出LFB(long-term feature banks)为视频模型提供长达60 s的时间支持,用于计算动作执行者之间的远程交互,可以获得很大的性能增益。Girdhar等人(2019)重新利用Transformer网络(Vaswani等, 2017),通过使用高分辨率的、特定于人的与类无关的查询,模型可以自动地学习每两个动作执行者之间的成对关系,并从他人的动作中提取出相关的语义上下文信息。AIA(asynchronous interaction aggregation network)(Tang等, 2020)通过多个类型的交互模块串行堆叠,将在前一个交互模块中得到增强的目标特征传递给后续模块,旨在对不同类型的交互作用进行融合,而不只是关注单一类型的交互。上述方法都是对动作执行者(actor)与周围上下文(context)之间直接的低阶交互关系进行建模,Pan等人(2021)提出显式建模间接的高阶交互关系,从而有效地推理人物在复杂场景中的行为,特别是在两个actor并非直接进行交互,而是通过context作为媒介产生关联的情况下。王东祺和赵旭(2022)构建了一个简洁有效的时序关联模块,借助门控循环单元建立了当前时刻与过去未来时刻的全局时序关联。

1.2 视频中的Attention机制

注意力机制(attention mechanism)最初起源于对人类的视觉研究,为了合理利用有限的视觉处理资源,人类会选择性地关注一部分信息,同时忽略其他可见信息。Transformer(Vaswani等, 2017)将注意力机制引入到机器翻译任务中,缓解了源序列与目标序列的长距离依赖问题。该模型是由一些堆叠的自注意力层(self-attention layer)和全连接层(fully-connected layer)组成的,自注意力模块通过关注所有位置并在嵌入空间中取它们的加权平均值来计算一个序列中某个位置的响应。Wang等人(2018)提出了Non-local模块,将注意力机制拓展为更通用的形式,使其能够在计算机视觉的诸多任务上应用。Non-local操作将某个位置的响应计算为输入特征映射中所有位置的加权和,以捕获长距离的依赖关系。这里的位置集合可以是空间、时间或时空,意味着该操作普遍适用于图像、序列和视频问题。LFB(long-term feature banks)(Wu等, 2019)也使用Non-local作为运算符,提取长期特征库内与当前动作相关的全局时序上下文信息。

1.3 图注意力网络

图注意力网络(graph attention network, GAT)(Velickovic等, 2018)常用于关系推理任务中,核心思想是在图算法中引入注意力机制,计算当前节点与邻居节点之间的权重系数,使得图网络能够更加关注重要的节点。算法流程如下:对于输入的所有节点,网络训练生成一个共享的权重矩阵,得到每个邻居节点的权重,然后对邻居节点的特征进行加权求和得到输出特征。它的头结构如图1所示,其中,邻接矩阵$\boldsymbol{A}$是一个二值矩阵,$\boldsymbol{A}_{ij}$代表节点i和节点j之间的邻接关系,若邻接则为1,不邻接则为0,通常邻接矩阵是预先定义好的,在后续特征更新时只需要考虑相邻的节点;相关性矩阵$\boldsymbol{E}$是对节点两两之间计算互相关系数得到的,$\boldsymbol{E}_{ij}$代表节点j对节点i的

重要性；而权重矩阵 $\boldsymbol{W}$ 由邻接矩阵 $\boldsymbol{A}$ 和相关性矩阵 $\boldsymbol{E}$ 进行点乘后，再通过 softmax 函数归一化得到，$\boldsymbol{W}_{ij}$ 代表节点 i 和节点 j 之间的权重系数。

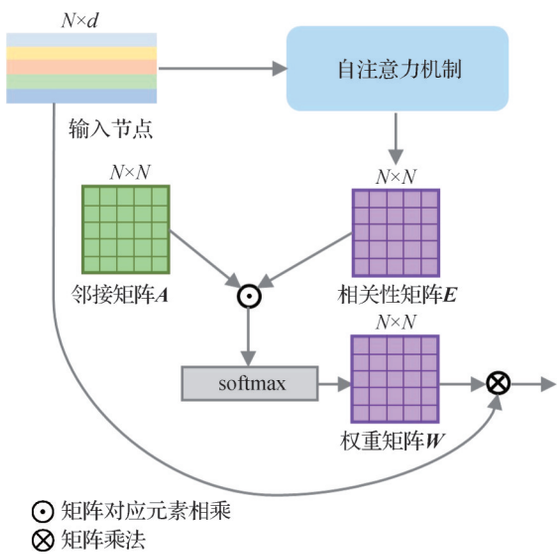


图 1　图注意力网络头结构

Fig. 1　Structure of graph attention heads

2　本文算法

本文的主要思想是通过建模动作执行者之间的交互作用来增强目标特征，提高交互类动作的识别性能，同时利用动作类别之间的依赖关系来增强类别表征，在一定程度上解决多标签分类问题。整体框架由 3 部分组成，分别是特征提取部分、长短期交互部分和分类器部分，如图 2 所示。其中，特征提取部分包括预训练好的人物检测器和骨干网络，长短期交互部分包括短期交互模块（short-term interaction module，STIM）和长期交互模块（long-term interaction module，LTIM），分类器部分设计类别关系模块（class relationship module，CRM）并采用双阶段分数融合（two-stage score fusion，TSSF）策略。本文算法的具体步骤如下：

1）在视频关键帧上使用微调过的 Faster R-CNN（region based convolutional neural network）网络（Ren 等，2017）对动作执行者进行目标检测，得到一系列检测框的位置坐标以及动作性分数，再利用预训练好的 SlowFast 骨干网络处理以关键帧为中心的视频片段，提取出检测框对应的区域特征；

2）STIM 模块采用两层图注意力网络 GAT，针对性地处理空间交互和短期内的时间交互，LTIM 模块引入长期特征库 LFB 为模型提供长期时间支持，计算动作执行者之间的远程交互，通过这两个模块对

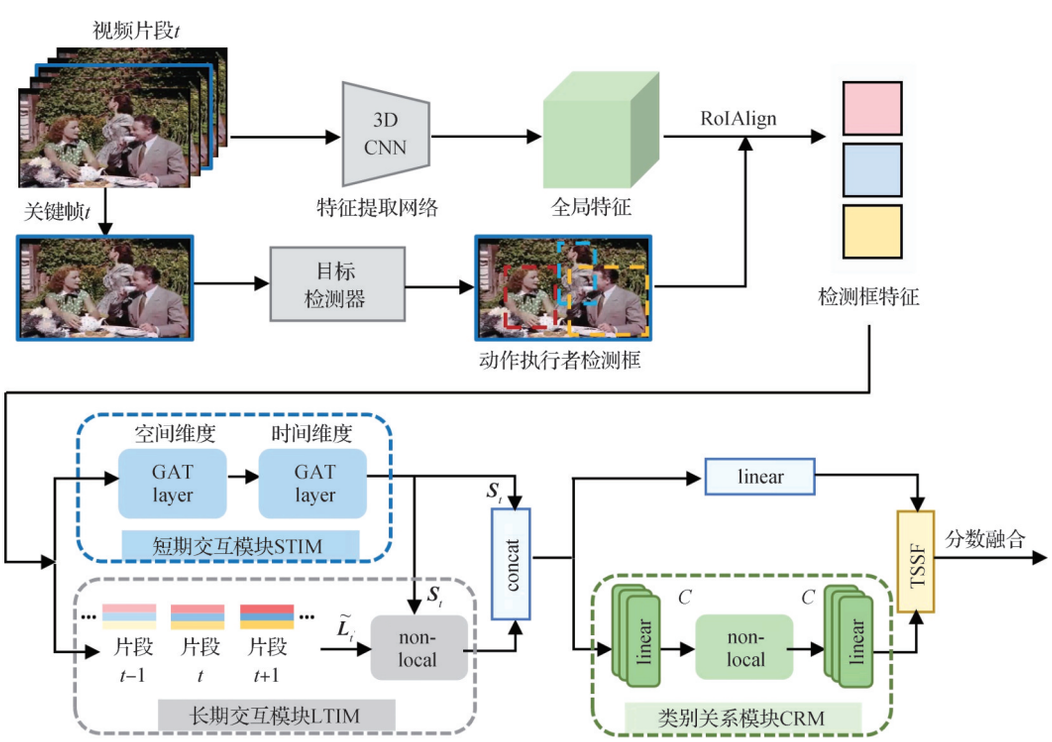


图 2　本文模型整体架构

Fig. 2　Overall architecture of the proposed model

动作执行者本身的区域特征进行增强，提高目标特征的表达能力；

3）设计CRM模块抽取每个动作类的表征，并通过自注意力机制计算它们之间的语义相关性对原始表征进行加权，得到增强后的各个类表征，最后根据不同模块的互补特性，将第1阶段的预测分数和新分数融合，更新每一类的概率得分。

2.1 特征提取部分

首先，本文采用在AVA v2.1数据集上微调过的Faster R-CNN网络作为人物检测器，对视频关键帧中出现的所有动作执行者进行目标检测，得到一系列检测框的位置坐标以及动作性分数，该检测器在AVA v2.1验证集上的mAP@IoU 0.5（mean average precision@intersection over union 0.5）可以达到93.9%。然后，使用预训练好的SlowFast骨干网络作为特征提取器，选取动作性分数大于阈值（这里为0.8）的检测框进行RoIAlign（region of interest align）（He等，2017），并执行最大值池化操作，得到2 304维的特征向量。最后，将所有检测框的位置坐标归一化到[0，1]区间，编码后加入对应的原始特征向量中，生成动作执行者的实例级表征，用于后续的动作分类过程。

2.2 长短期交互部分

2.2.1 短期交互模块STIM

STIM模块旨在建模短期时间内动作执行者两两之间的交互关系。考虑到不同的动作执行者在同一时刻执行的动作具有交互性，而同一个动作执行者在一段短期时间内所执行的动作具有连续性，且空间交互和时间交互本质上是不同的，本文对空间维度和时间维度上的交互作用分别进行建模。如图3所示，STIM模块采用解耦机制，通过两层图注意力网络针对性地处理空间交互作用和时间交互作用，同时也降低了网络的学习难度。

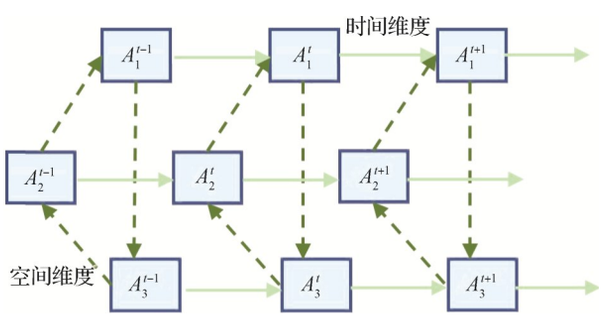
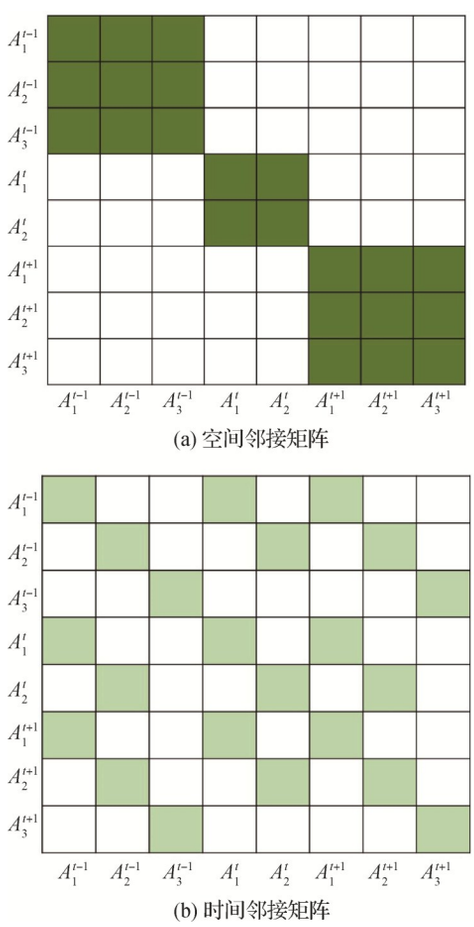


图3 时空解耦机制

Fig. 3 Spatial-temporal decoupling mechanism

图注意力网络GAT的头结构如图1所示。显然，可以通过改变邻接矩阵$\boldsymbol{A}$的值来调整两两节点之间的邻接关系。例如，对于节点i，若节点j与之相邻，则$\boldsymbol{A}_{ij}$的值为1，否则为0。基于上述分析，在考虑空间维度上的交互作用时，需要将同一帧内不同动作执行者的节点两两邻接，此时邻接矩阵的分布如图4（a）所示。此外，为了引入不同动作执行者的空间相对位置关系，本文首先计算两两之间的欧氏距离，并用softmax函数归一化处理，形成一个和邻接矩阵$\boldsymbol{A}$形状相同的距离矩阵，最后将其与$\boldsymbol{A}$进行点乘，作为新的邻接矩阵。

类似地，在考虑时间维度上的交互作用时，由于同一个动作执行者在一段短期时间内所执行的动作具有连续性，所以模型将同一个动作执行者在不同时间步的节点两两邻接，此时邻接矩阵的分布如图4（b）所示，其中A_k^t表示第t帧的第k个动作执行者



(a) 空间邻接矩阵

(b) 时间邻接矩阵

图4 邻接矩阵分布示例

Fig. 4 Examples of adjacency matrix distribution

((a) spatial adjacency matrix; (b) temporal adjacency matrix)

节点。本文使用Singh等人(2017)提出的贪心算法对同一个动作执行者的检测框进行跟踪，采取的贪心策略是匹配时选择下一帧中与当前框交并比(intersection over union，IoU)最大的检测框，且IoU要大于预设阈值(通常为0.5)。

GAT的特征更新规则具体为：对于输入节点i，假设它的特征向量为$\boldsymbol{h}_i$，邻接节点集合为$\boldsymbol{N}_i$，则经过GAT之后的节点特征更新为

$$\boldsymbol{h}_i' = \sigma\left(\sum_{j\in \boldsymbol{N}_i} \boldsymbol{W}_{ij}\boldsymbol{h}_j\right) \tag{1}$$

$$\boldsymbol{W}_{ij} = \mathrm{softmax}\left(\sigma\left(\boldsymbol{w}_b^{\mathrm{T}}\left[\boldsymbol{h}_i \| \boldsymbol{h}_j\right]\right)\right) \tag{2}$$

式中，σ是线性整流函数ReLU(rectified linear unit)，$\boldsymbol{w}_b$是一个可学的向量。

2.2.2　长期交互模块LTIM

LTIM模块旨在提供时序上的长期支持信息，从而帮助模型更好地理解持续时间长且内容复杂的视频，更准确地推断动作执行者在当前时刻的行为状态。如图5所示，对于目标帧t中发生的listen(e.g.，to music)动作，如果只考虑短期内时序上与t相邻的前后几帧，将无法确切地得出动作执行者正在listen(e.g.，to music)这个结论。然而，当接收到更长时间范围的信息之后，模型可以自动地捕捉与目标帧中动作相关的时空上下文，大幅提高了模型的置信度。

图5　长期特征库的构建

Fig. 5　Construction of long-term feature banks

本文在LTIM模块的设计上遵循LFB(long-term feature banks)(Wu等，2019)的思想，使用滑动窗口灵活地构建包含历史和未来时刻的长期特征库，缓解了由于GPU(graphics processing unit)内存限制导致视频输入太短的问题。假设当前时刻为t，则第t帧的特征矩阵可表示为$\boldsymbol{L}_t \in \mathbf{R}^{N_t\times d}$，包含当前帧内共$N_t$个动作执行者的特征向量，每个特征向量的维数为$d$。那么第$t$帧的长期特征库可表示为$\tilde{\boldsymbol{L}}_t = [\boldsymbol{L}_{t-w},\cdots,\boldsymbol{L}_{t+w}]$，即以当前帧$t$为中心，前后各取$w$帧的特征矩阵组成，则有$\tilde{\boldsymbol{L}}_t \in \mathbf{R}^{N\times d}$，其中$N = \sum_{t'=t-w}^{t+w} N_{t'}$代表$2w+1$长度的特征库内所有动作执行者的数量。

LTIM模块的算法流程如图6(a)所示。将第t帧内目标经过STIM模块增强后的特征记做$\boldsymbol{S}_t$，再将目标特征$\boldsymbol{S}_t$和对应的长期特征库$\tilde{\boldsymbol{L}}_t$通过算子(operator)进行作用，提取出与之相关的全局时序上下文信息。这里采用的算子是改造后的non-local模块，具体操作为：首先对$\boldsymbol{S}_t$和$\tilde{\boldsymbol{L}}_t$进行线性变换降至d维，将$\boldsymbol{S}_t$作为query，$\tilde{\boldsymbol{L}}_t$作为key-value对，在嵌入空间中使用内积计算它们之间的语义相关性。然后对计算结果进行尺度缩放(除以$\sqrt{d}$)并通过softmax函数归一化，尺度缩放的目的在于减小点积值的数量级，避免后续的softmax函数发生梯度消失问题。接着将归一化之后的系数作为权重，对value加权求和，再利用LayerNorm函数使得每个样本内的分布一致，最后提取出时序上的全局信息并更新下一层的输入特征。本文使用了2层non-local的叠加结构(NL)，即

$$\boldsymbol{S}_t' = NL_{\theta_1}(\boldsymbol{S}_t,\tilde{\boldsymbol{L}}_t) \tag{3}$$

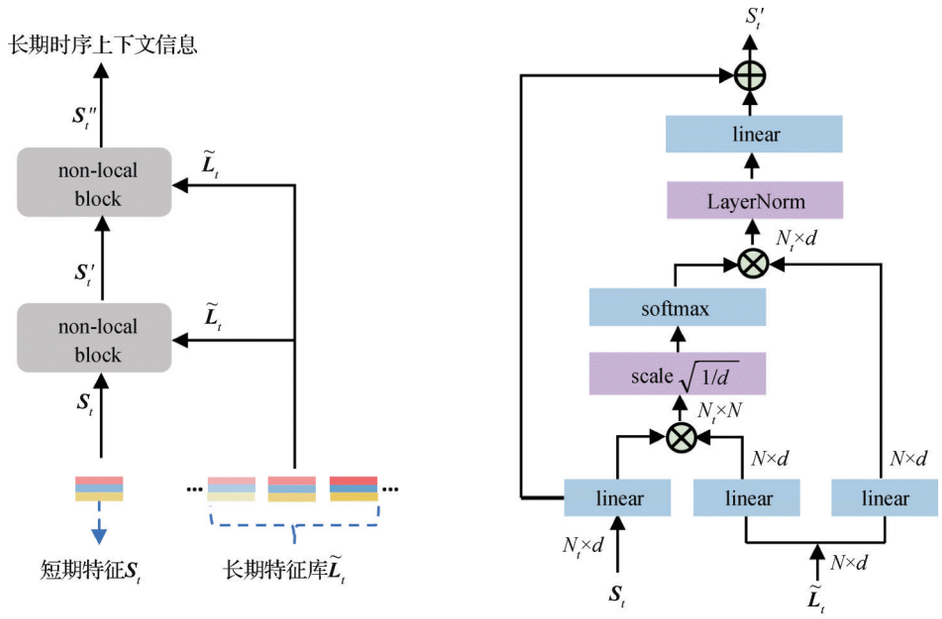
$$\boldsymbol{S}_t'' = NL_{\theta_2}(\boldsymbol{S}_t',\tilde{\boldsymbol{L}}_t) \tag{4}$$

式中，θ_1和θ_2是可学参数，non-local的结构如图6(b)所示。

2.3　分类器部分

2.3.1　类别关系模块CRM

针对多标签分类问题，现有方法普遍采用二元交叉熵(binary cross entropy loss，BCELoss)组合作为模型的代价函数，操作简单且效果较好，因而获得了广泛的应用。但是这种实现方式意味着将多标签分类问题视做多个类上的二分类问题，最终预测时各个类别是独立进行的，忽略了类别之间的内在依赖关系。因此，本文在AVA v2.1训练集上对样本的标签共现信息进行了统计，这里的统计指标为在动作i发生的条件下动作j发生的概率$p(j|i)$。结果如图7



(a) 长期交互模块流程图　　(b) non-local模块组成

图6　LTIM模块结构

Fig. 6　Structure of LTIM module ((a) flow chart of long-term interaction module; (b) composition of non-local block)

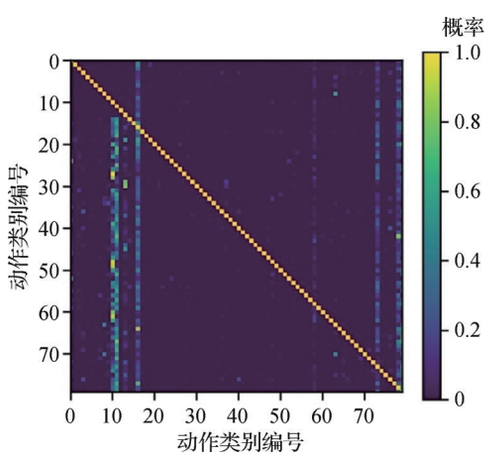


图7　训练样本的标签共现信息

Fig. 7　Label co-occurrence information of training samples

所示，其中颜色越亮的位置表示对应动作对的共现概率越高，横纵坐标轴为该数据集的所有动作类别，此处用数字编号的形式表示。

特别地，本文定义了3种动作类之间的依赖关系。1)$p(j|i) \approx 1$，说明给定动作j和条件动作i极有可能同时发生，它们之间存在先决条件关系；2)$p(j|i) \approx 0$，说明给定动作j和条件动作i极不可能同时发生，它们之间存在排除关系；3)$p(j|i) \in (0, 1)$，说明给定动作j和条件动作i有一定的概率同时发生，它们之间存在重叠关系。

CRM模块的目的是使网络自动地学习动作执行者在同一时刻不同动作类之间的语义相关性，自适应地增强各个动作类的表征，从而提升多标签分类的性能。尤其是对正样本数量比较少的动作类而言，网络可能很难学习到有效且鲁棒的类别表征，此时可以利用与之关联性较强或者共现性较高的其他动作类别，通过特征融合的方式增强自身的表征，提高识别准确率。该模块的结构设计如图8所示，首先对动作执行者的d维特征向量进行C次线性变换(C为动作类别数)，分别提取各个动作类别的表征；然后通过自注意力机制建模不同动作类别之间的语义相关性。具体来说，首先对类别特征进行线性变换分别得到对应的query，key和value，并分别记为Q_N，K_N，和V_N。然后，计算Q_N与K_N之间的内积得到类别之间的类别依赖矩阵。再之后，利用该相关性矩阵对V_N进行加权，得到增强后的类别特征。最后，将增强后的特征输入到分类器中，预测出每个动作类的概率得分。

2.3.2　双阶段分数融合策略TSSF

分类器部分设计了CRM模块，利用不同动作类别之间的依赖关系对特征进行增强，以达到提升多标签分类性能的目的。该模块对动作执行者进行了二次动作分类，对于正样本数量少的动作类别，网络可以更容易地学习到其表征，弥补了第1阶段特征不够好的缺陷。

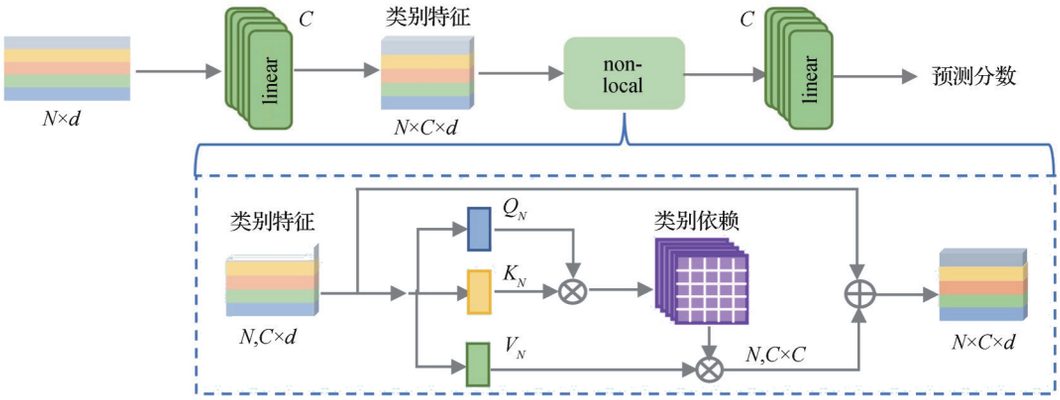


图 8　CRM 模块结构

Fig. 8　Structure of CRM module

由于第 1 阶段对少数类样本的区分能力不强，预测分数具有保守性，而 CRM 模块在第 2 阶段中针对性地改善了该问题，所以提出双阶段分数融合策略对每个动作类的预测分数进行融合，可以采用取最小值、最大值或者平均值的方式得到最终的概率分数，具体为

$$s_c^* = \begin{cases} \min(s_c^1, s_c^2) \\ \max(s_c^1, s_c^2) \\ \text{avg}(s_c^1, s_c^2) \end{cases} \tag{5}$$

式中，s_c^1 和 s_c^2 分别表示第 1 阶段和第 2 阶段动作类 c 的预测分数，s_c^* 表示融合之后的概率得分。

本文中，第 1 阶段长短期交互部分和第 2 阶段分类器部分是联合训练的，模型的总体代价函数为

$$L = L_1 + L_2 \tag{6}$$

$$L_1 = -\frac{1}{N}\sum_i^N\sum_j^C [\, y_j^{(i)} \log(s_j^{1(i)}) + (1 - y_j^{(i)}) \log(1 - s_j^{1(i)})\,] \tag{7}$$

$$L_2 = -\frac{1}{N}\sum_i^N\sum_j^C [\, y_j^{(i)} \log(s_j^{2(i)}) + (1 - y_j^{(i)}) \log(1 - s_j^{2(i)})\,] \tag{8}$$

式中，L_1 和 L_2 分别为第 1 阶段和第 2 阶段的代价函数，L 为这两个阶段的总代价函数，$y_j^{(i)}$ 为第 i 个样本的第 j 个动作类上的标签，$s_j^{1(i)}$ 和 $s_j^{2(i)}$ 分别为第 1 阶段和第 2 阶段在第 i 个样本的第 j 个动作类上的预测分数。

3　实验分析

3.1　数据集

本文在 AVA v2.1 数据集上进行实验，该数据集由从 YouTube 中收集的 430 个视频组成，其中训练集包含 235 个，验证集包含 64 个，测试集包含 131 个，所占比例分别为 55%、15% 和 30%。对于每个视频，选取 15 min 的固定长度进行标注，采样频率为 1 Hz，即每个视频包含 900 个关键帧。标注时，对关键帧中出现的所有动作执行者，逐一确定其空间位置并选择适当数量的动作标签。该数据集的动作标签共 80 种，按动作特点分为 3 类。第 1 类为人体姿势动作，包含 14 个子类；第 2 类为人—物交互动作，包含 49 个子类；第 3 类为人—人交互动作，包含 17 个子类。每个动作实例最多可分配 7 个动作标签，包括 1 个人体姿势标签、0～3 个人—物交互标签和 0～3 个人—人交互标签。综上，AVA v2.1 数据集共包含 211 k 个训练实例、57 k 个验证实例和 117 k 个测试实例，遵循主流方法的设置，本文在训练集上进行模型训练，并在验证集上评估结果。

3.2　评价指标

本文采用阈值取 0.5 时的帧级均值平均精度（frame-level mAP@IoU 0.5）作为评价指标，在最终的结果评估时，根据统一的准则，选取了验证集实例数不少于 25 的 60 个动作类，忽略了验证集实例数少于 25 的剩余 20 个动作类。

3.3　模型设置

实验基于深度学习框架 pytorch 1.6.0 部署，模型的训练和测试过程均基于 pytorch 完成。本文首先以 30 帧/s 的帧率对视频进行抽帧，并将每秒的中间帧定义为关键帧；然后使用微调过的 Faster R-CNN 网络在关键帧上对动作执行者进行目标检测，得到一系列检测框的位置坐标和动作性得分；最后利用预训练好的 SlowFast 骨干网络提取每个检测

框对应的区域特征，训练时选取动作性得分不低于0.9的检测框以及ground-truth框，测试时仅使用得分不低于0.8的检测框。另外，实验在NVIDIA Geforce GTX 1080Ti（11 GB内存）上完成，模型采用Adam（adaptive moment estimation）优化器加速训练，初始学习率设置为0.000 01，每个训练批次包含16组数据，完成50个训练周期。

3.4 实验结果

为了验证本文模型中各模块的有效性，进行了一系列消融实验，这里以SlowFast为基准模型，具体为：1）添加STIM模块；2）添加STIM模块，分类器部分添加CRM模块并采用TSSF策略；3）添加STIM模块和LTIM模块；4）添加STIM模块和LTIM模块，分类器部分添加CRM模块并采用TSSF策略。同时在AVA v2.1数据集上进行定量和定性分析。然后将本文方法与其他视频动作检测方法进行对比，验证本文方法的先进性。

3.4.1 定量分析

首先对本文模型进行消融实验，分析模型的各个组成部分对最终性能的影响，结果如表1所示。然后为了更深入地分析模型，分别评估了各模块在人体姿势类别、人—物交互类别和人—人交互类别上的性能，结果如表2所示。此外，表3中列出了各参数不同取值的组合，并计算每种组合下本文模型在验证集上的指标。最后在表4中给出了各个模块的参数量和推理阶段每个关键帧的检测时间。

表1 不同的模块组合在AVA v2.1验证集上的性能

Table 1 Performance of different module combinations on the validation set of AVA v2.1

方法	mAP/%
SlowFast	28.2
SlowFast + STIM	29.8
SlowFast + STIM + CRM	30.1
SlowFast + STIM + LTIM	30.6
SlowFast + STIM + LTIM + CRM（本文）	**31.0**

注：加粗字体表示最优结果。

表2 人体姿势、人—物交互和人—人交互类别的消融分析（mAP）

Table 2 Ablation analysis on Human Pose, Human-Object Interaction and Human-Human Interaction categories in terms of mAP

/%

方法	人体姿势	人—物交互	人—人交互
SlowFast	48.8	19.7	28.9
SlowFast + STIM	49.9	20.9	31.6
SlowFast + STIM + CRM	50.1	21.2	31.8
SlowFast + STIM + LTIM	50.5	22.1	31.8
SlowFast + STIM + LTIM + CRM（本文）	**50.8**	**22.3**	**32.5**

注：加粗字体表示各列最优结果。

表1展示了模型各组成部分对最终性能的影响。可以看出，STIM模块利用图注意力网络显式建模动作执行者之间的短期交互关系，以聚合与动作相关的局部时空信息，有助于模型更好地学习表征，平均精度均值（mean average precision，mAP）比基准模型的结果提升了1.6%。LTIM模块通过提供长期的时间支持信息，计算动作执行者之间的远程依赖，有助于更精确地推断当前时刻的动作状态，兼顾了时序上的全局信息利用，mAP在STIM模块的基础上提升了0.8%。而分类器部分的CRM模块与TSSF策略结合，利用动作类别之间的标签依赖关系增强特征，同时针对不同模块的互补性进行分数融合。CRM模块具有通用性，无论是添加到STIM模块上（+0.3% mAP），还是添加到最终的模型上（+0.4% mAP），都能在一定程度上解决多标签分类问题，提升模型的准确率。

除了给出模型在60个测试子类上的均值平均精度外，本文分别评估了各模块在人体姿势、人—物交互和人—人交互类别上的结果，如表2所示。与

基准相比，最终模型在人体姿势、人—物交互和人—人交互类别上的均值平均精度分别增加了 2.0%，2.6%，3.6%，这表明本文方法对建模交互作用和类别依赖都是有效的。同时，本文计算了各个动作类的指标用于对比，如图 9 所示。可以看到，本文模型在 sing to（e. g. ，self，a person，a group）、play musical instrument、cut 等动作类上均有明显的性能改进，可能是因为这些动作类的识别需要对动作执行者之间的交互关系进行推理，或者动作本身具有很强的长期时序依赖性。

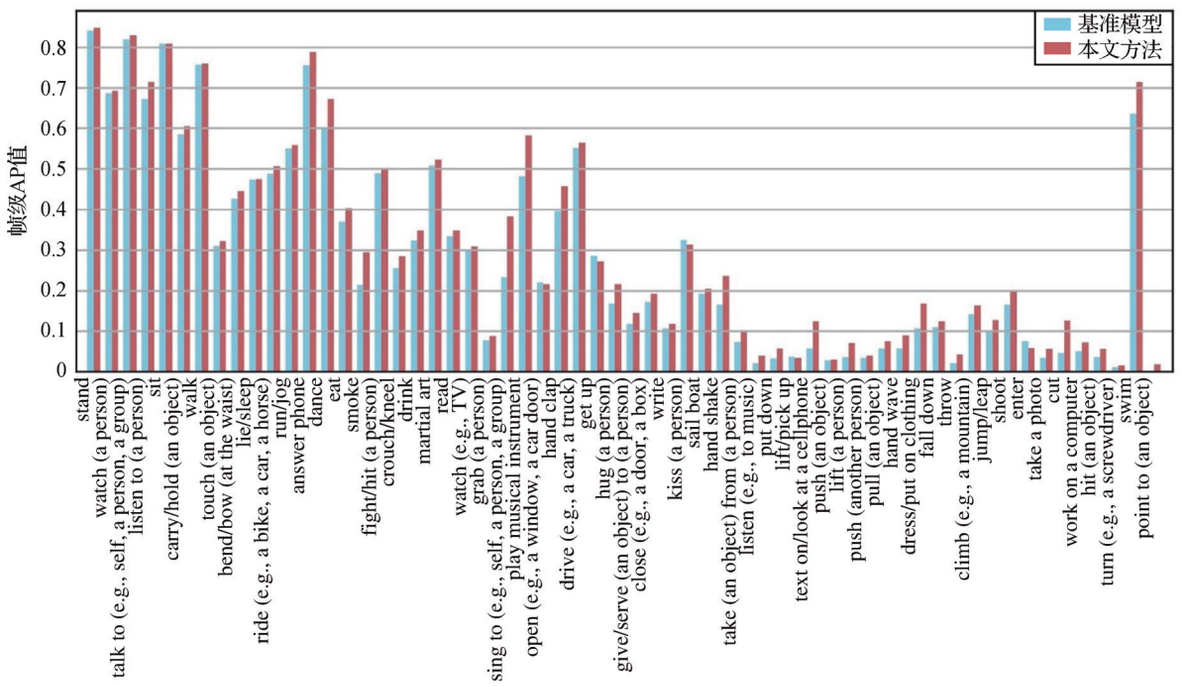


图 9　在 AVA v2. 1 验证集上本文模型和基准模型的各类别指标

Fig. 9　Per-category results for the proposed model and the baseline on the validation set of AVA v2. 1

表 3 展示了不同参数组合对模型性能的影响。这里 Heads 指的是 STIM 模块中图注意力网络 GAT 的头数目，而 Dim 和 Layers 分别指的是 CRM 模块中类别表征的维数以及注意力层的数目。不难发现，随着 GAT 头数目的增加，模型性能进一步提升，这是因为在注意力机制中采用多头的设置可以将输入映射到多个不同的语义空间中，使模型获取的信息更为丰富，也能起到集成的作用，在一定程度上降低了网络发生过拟合的风险。而 CRM 模块中类别表征的维数并不是越大越好，过大的特征维数会包含许多冗余信息，也会增加不必要的模型复杂度。寻找最优的参数配置实际上就是追求一种模型复杂度和模型性能之间的平衡。本文最终采用的设置是取 Heads = 8，Dim = 32，Layers = 3，对所有的实验统一。此外，本文对模型的参数量和推理速度进行了分析。表 4 中的数据体现了各模块之间存在的显著差异，在参数量方面，LTIM < CRM < STIM；在检测时间方面，STIM < LTIM < CRM。所以，在设计模块时

表 3　考虑不同参数组合时 AVA v2. 1 验证集上的结果

Table 3　Results on the validation set of AVA v2. 1 considering different parameter combinations

参数组合		mAP/%
Heads = 2	Dim = 32, Layers = 2	30.3
	Dim = 32, Layers = 3	30.4
	Dim = 64, Layers = 2	30.6
	Dim = 64, Layers = 3	30.5
Heads = 4	Dim = 32, Layers = 2	30.9
	Dim = 32, Layers = 3	30.7
	Dim = 64, Layers = 2	30.6
	Dim = 64, Layers = 3	30.5
Heads = 8	Dim = 32, Layers = 2	30.8
	Dim = 32, Layers = 3	**31.0**
	Dim = 64, Layers = 2	30.6
	Dim = 64, Layers = 3	30.7

注：加粗字体表示最优结果。

表4　本文模型的参数量和推理速度

Table 4　The parameters and inference speed of our model

方法	参数量/M	检测时间/ms
本文 w/o STIM	19.20	2.09
本文 w/o LTIM	35.95	1.57
本文 w/o CRM	26.00	1.53
本文	40.46	2.82

注：w/o代表without。

不能仅参考某一项指标的值，而应该综合考虑，优化模型的整体性能。

3.4.2　定性分析

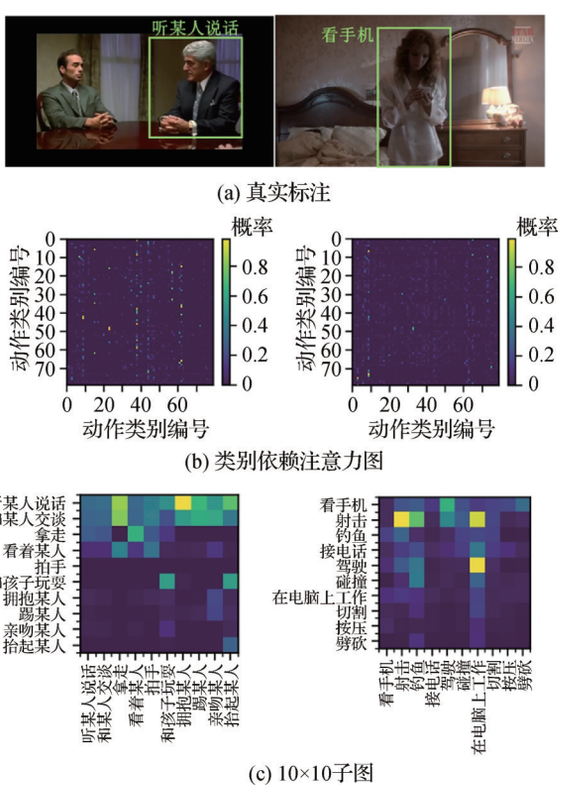
为了定性地评估本文模型，将视频关键帧的动作检测结果可视化，对几个具有挑战性的示例进行性能比较，如图10所示。在这些示例中，动作参与者所执行的动作具有长期时序依赖性并且涉及到与其他对象的交互。图10(a)(b)分别为基准模型和本文模型的检测结果。第1行的动作参与者正在执行sing to(e. g. , self, a person, a group)这个动作，该动作的建模不仅需要考虑交互对象(即听众)，还需要整合时域上的全局信息。本文模型针对动作执行者特别融合了交互对象的表征，并且精准捕捉了与动作相关的长期时序上下文信息，在该样例中实现了准确的预测。而基准模型只利用到动作执行者本身的区域特征，也没有考虑到动作在时序上的全局依赖性，很容易造成误判。从第2行可以看到一个男人正在切割木板，这同样也需要利用时序上的长期特征来辅助动作推断。第3行展示的是两个人打架的场景，建模二者之间的交互作用显然对模型性能的提升大有裨益。

本文对CRM模块中间过程的类别依赖注意力图进行了可视化，如图11所示。图11(a)为选取的两个动作示例及其真实标注。第1个示例的动作执行者标签为“听某人说话”，第2个示例为“看手机”。图11(b)为生成的类别依赖注意力图，颜色越亮的位置表示响应越强，即对应的两个动作类之间的依赖性高，颜色越暗的位置表示响应越弱，即对应的两个动作类之间的依赖性低。图11(c)为进一步提取的10×10子注意力图，其中与当前动作相关的动作往往具有较强的响应。例如“听某人说话”、“拿走”和“看着某人”通常与“和某人交谈”同时发生，而不相关的动作普遍响应较弱，例如“拍手”和“抬起某

(a) 基准模型　　(b) 本文模型

图10　基准模型和本文模型在AVA v2.1验证集的可视化结果

Fig. 10　Visualization results for the proposed model and the baseline on the validation set of AVA v2.1 ((a) baseline model; (b) ours)



(a) 真实标注

(b) 类别依赖注意力图

(c) 10×10子图

图11　在AVA v2.1验证集上类别依赖注意力图的可视化示例

Fig. 11　Visualization examples of class dependency attention maps on the validation set of AVA v2.1 ((a) ground-truth; (b) class dependency attention map; (c) 10 × 10 subset of the attention map)

人”没有明显的关联关系，说明了CRM模块对类别依赖建模的合理性和可靠性。

3.4.3　与其他方法的性能比较

为了更全面地验证本文方法的性能，与ACRN（actor-centric relation network）（Sun等，2018）、VAT（video action transformer network）（Girdhar等，2019）、LFB（Wu等，2019）、Context-Aware RCNN（Wu等，2020）、SlowFast（Feichtenhofer等，2019）、ACAR-Net（actor-context-actor relation network）（Pan等，2021）、LSTC（long-short term context）（Li等，2021）和IGMN（identity-aware graph memory network）（Ni等，2021）等方法在AVA v2.1数据集上进行比较，实验结果如表5所示。可以看出，本文方法的mAP@IoU 0.5指标分别提高了13.6%，6.0%，3.6%，3.0%，2.8%，1.0%，1.0%，0.8%。

表5　不同方法在AVA v2.1验证集上的实验结果

Table 5　Experimental results of different methods on the validation set of AVA v2.1

模型	mAP/%
ACRN（Sun等，2018）	17.4
VAT（Girdhar等，2019）	25.0
LFB（Wu等，2019）	27.4
Context-Aware RCNN（Wu等，2020）	28.0
SlowFast（Feichtenhofer等，2019）	28.2
ACAR-Net（Pan等，2021）	30.0
LSTC（Li等，2021）	30.0
IGMN（Ni等，2021）	30.2
本文	**31.0**

注：加粗字体表示最优结果。

ACRN通过网络自动挖掘场景中与动作执行者相关的时空元素，生成用于动作分类的关系特征，但没有显式地对实体之间的交互作用进行建模。VAT对Transformer网络进行改造并应用于视频动作检测任务，但同样也是隐式地提取与动作执行者相关的周围上下文信息。LFB为模型提供长期的时间支持，通过计算实体之间的远程交互来建模全局时序依赖关系，但没有强调利用动作在时序上的局部关联性。Context-Aware RCNN在提取动作执行者的区域特征之前，对周围图像块进行裁剪和调整，避免空间细节的损失，但在融合上下文信息方面缺少相应设计，只是简单地进行特征拼接。ACAR-Net提出了高阶交互关系的概念，除了动作执行者之间直接的一阶关系之外，还考虑了间接的二阶交互关系，但没有针对性地处理这两种关系。LSTC利用视频信号之间的时间依赖性进行动作识别，将其分解为短期依赖和长期依赖独立推断，但同样忽视了空间维度和时间维度的异质性。此外，上述方法均没有考虑AVA v2.1数据集本身存在的多标签问题。本文以时空解耦的方式显式建模动作执行者之间的短期交互作用，同时引入长期特征库兼顾了时序上的局部信息和全局信息，在分类器部分利用类别依赖关系增强表征，且根据不同模块的互补性分阶段预测并对分数进行融合，显著提升了动作检测的效果。

4　结　论

本文提出了一种同时考虑交互关系和类别依赖的视频动作检测方法，试图去建模动作执行者之间的交互作用以及动作类别之间的语义相关性。与对比算法相比，本文在建模交互关系时充分利用了视频信号独有的时空特性，以时空解耦的方式显式建模动作执行者之间的短期交互作用，并引入长期特征库计算远程交互以提取长期时序上下文信息，兼顾了时序上的局部关联和全局依赖。此外，在处理多标签数据集时，本文进行了针对性的设计，通过类别关系模块提取特定于类别的表征，并使用注意力机制计算动作类之间的语义相关性以增强表征。最后根据不同模块的互补特性，提出双阶段分数融合策略更新最终的概率得分。本文方法在一定程度上克服了对比算法存在的局限性，在AVA v2.1数据集上的定量和定性分析结果表明了本文方法的有效性和鲁棒性。相较于其他的视频动作检测方法，本文方法大幅提升了交互类动作的识别效果，同时以较低的计算代价解决了多标签问题。后期研究将进一步优化网络结构，例如改进目标特征与长期特征库的作用机制，以使mAP指标得到更显著的提升。

参考文献（References）

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,

USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Feichtenhofer C, Fan H Q, Malik J and He K M. 2019. SlowFast networks for video recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6201-6210 [DOI: 10.1109/ICCV.2019.00630]

Girdhar R, Carreira J J, Doersch C and Zisserman A. 2019. Video action transformer network//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 244-253 [DOI: 10.1109/CVPR.2019.00033]

Gu C H, Sun C, Ross D A, Vondrick C, Pantofaru C, Li Y Q, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, Schmid C and Malik J. 2018. AVA: a video dataset of spatio-temporally localized atomic visual actions//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6047-6056 [DOI: 10.1109/CVPR.2018.00633]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397 [DOI: 10.1109/TPAMI.2018.2844175]

Li Y X, Zhang B S, Li J, Wang Y B, Lin W Y, Wang C J, Li J L and Huang F Y. 2021. LSTC: boosting atomic action detection with long-short-term context//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 2158-2166 [DOI: 10.1145/3474085.3475374]

Luo H L, Tong K and Kong F S. 2019. The progress of human action recognition in videos based on deep learning: a review. Acta Electronica Sinica, 47(5): 1162-1173 (罗会兰, 童康, 孔繁胜. 2019. 基于深度学习的视频中人体动作识别进展综述. 电子学报, 47(5): 1162-1173) [DOI: 10.3969/j.issn.0372-2112.2019.05.025]

Ni J C, Qin J and Huang D. 2021. Identity-aware graph memory network for action detection//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 3437-3445 [DOI: 10.1145/3474085.3475503]

Pan J T, Chen S Y, Shou M Z, Liu Y, Shao J and Li H S. 2021. Actor-context-actor relation network for spatio-temporal action localization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 464-474 [DOI: 10.1109/CVPR46437.2021.00053]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Singh G, Saha S, Sapienza M, Torr P and Cuzzolin F. 2017. Online real-time multiple spatiotemporal action localisation and prediction//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3657-3666 [DOI: 10.1109/ICCV.2017.393]

Sun C, Shrivastava A, Vondrick C, Murphy K, Sukthankar R and Schmid C. 2018. Actor-centric relation network//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 335-351 [DOI: 10.1007/978-3-030-01252-6_20]

Tang J J, Xia J, Mu X Z, Pang B and Lu C W. 2020. Asynchronous interaction aggregation for action detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 71-87 [DOI: 10.1007/978-3-030-58555-6_5]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Velickovic P, Cucurull G, Casanova A, Romero A, Liò P and Bengio Y. 2018. Graph attention networks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR

Wang D Q and Zhao X. 2022. Class-aware network with global temporal relations for video action detection. Journal of Image and Graphics, 27(12): 3566-3580 (王东祺, 赵旭. 2022. 类别敏感的全局时序关联视频动作检测. 中国图象图形学报, 27(12): 3566-3580) [DOI: 10.11834/jig.211096]

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Wu C Y, Feichtenhofer C, Fan H Q, He K M, Krahenbuhl P and Girshick R. 2019. Long-term feature banks for detailed video understanding//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 284-293 [DOI: 10.1109/CVPR.2019.00037]

Wu J C, Kuang Z H, Wang L M, Zhang W and Wu G S. 2020. Context-Aware RCNN: a baseline for action detection in videos//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 440-456 [DOI: 10.1007/978-3-030-58595-2_27]

作者简介

贺楚景,女,硕士研究生,主要研究方向为计算机视觉。
E-mail: hcj1998@mail.ustc.edu.cn
王子磊,通信作者,男,副教授,博士生导师,主要研究方向为计算机视觉和模式识别。E-mail: zlwang@ustc.edu.cn
刘钦颖,男,博士研究生,主要研究方向为计算机视觉。
E-mail:lydyc@mail.ustc.edu.cn