E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)02-0406-15

论文引用格式:Cheng Y, Gao Y Y, Wang J, Yang L, Xu X L, Cheng Y and Zhang K H. 2025. Combining dilated convolution and multiscale fusion temporal action detection. Journal of Image and Graphics, 30(02):0406-0420(程勇, 高园元, 王军, 杨玲, 许小龙, 程遥, 张开华. 2025. 结合扩张卷积与多尺度融合的实时时空动作检测. 中国图象图形学报, 30(02):0406-0420)[DOI:10.11834/jig.240098]

结合扩张卷积与多尺度融合的实时时空动作检测

程勇[1*],高园元[1],王军[1],杨玲[1],许小龙[1],程遥[1],张开华[2]

1. 南京信息工程大学软件学院,南京　210044; 2. 南京信息工程大学计算机学院,南京　210044

摘　要:**目的**　时空动作检测任务旨在预测视频片段中所有动作的时空位置及对应类别。然而,现有方法大多关注行动者的视觉和动作特征,忽视与行动者交互的全局上下文信息。针对当前方法的不足,提出一种结合扩张卷积与多尺度融合的高效时空动作检测模型(efficient action detector,EAD)。**方法**　首先,利用轻量级双分支网络同时建模关键帧的静态信息和视频片段的动态时空信息。其次,利用分组思想构建轻量空间扩张增强模块提取全局性的上下文信息。然后,构建多种DO-Conv结构组成的多尺度特征融合单元,实现多尺度特征捕获与融合。最后,将不同层次的特征分别送入预测头中进行检测。**结果**　实验在数据集UCF101-24和AVA(atomic visual actions)中进行,分析了EAD与现有算法之间的检测对比结果。在UCF101-24数据集上的帧平均准确度(frame-mAP)和视频平均准确度(video-mAP)分别为80.93%和50.41%,对于基线方法的漏检、错检现象有所改善;在AVA数据集上的frame-mAP达到15.92%,同时保持较低的计算开销。**结论**　通过与基线及目前主流方法比较,EAD以较低的计算成本建模全局关键信息,提高了实时动作检测准确度。

关键词:深度学习;时空动作检测(STAD);双分支网络;扩张增强模块(DAM);多尺度融合

Combining dilated convolution and multiscale fusion temporal action detection

Cheng Yong[1*], Gao Yuanyuan[1], Wang Jun[1], Yang Ling[1], Xu Xiaolong[1], Cheng Yao[1], Zhang Kaihua[2]

1. School of Software, Nanjing University of Information and Science Technology, Nanjing　210044, China;

2. School of Computer Science, Nanjing University of Information and Science Technology, Nanjing　210044, China

Abstract: **Objective** Spatial-temporal action detection (STAD) represents a significant challenge in the field of video understanding. The objective is to identify the temporal and spatial localization of actions occurring in a video and categorize related action classes. The majority of existing methods rely on the backbone for the feature modeling of video clips, which captures only local features and ignores the global contextual information of the interaction with the actors. The results are represented by a model that cannot fully comprehend the nuances of the entire scene. The current mainstream methods for real-time STAD tasks are dual-stream network-based methods. However, a simple channel-by-channel connection is typically employed to handle dual-branch network fusion, which results in a significant redundancy of the fused features and certain semantic differences in the branch features. This scheme affects the accuracy of the model. Here, an efficient STAD model called the efficient action detector (EAD), which can address the shortcomings of current methods, is

收稿日期:2024-02-22;**修回日期**:2024-06-03;**预印本日期**:2024-06-10

*通信作者:程勇　yongcheng@nuist.edu.cn

基金项目:国家自然科学基金项目(41975183, 41875184)

Supported by: National Natural Science Foundation of China (41975183, 41875184)

proposed. **Method** The EAD model consists of three key components: the 2D branch, the 3D branch, and the fusion head. Among them, the 2D branch consists of a pretrained 2D backbone network, feature pyramid, and decoupling head; the 3D branch consists of a 3D backbone network and augmentation module; and the fusion head consists of a multiscale feature fusion unit (MSFFU) and a prediction head. First, key frames are extracted from the video clips and fed into the pretrained 2D branch backbone (YOLOv7) to detect the actors in the scene and obtain spatially decoupled features, which are classification features and localization features. Video spatial-temporal features are extracted from video clips via a pretrained lightweight video backbone network (Shufflenetv2). Second, the lightweight spatial dilated augmented module (LSDAM) uses the grouping idea to address spatial-temporal features, which serves to save resources. LSDAM consists of a dilated module (DM) and a spatial augmented module (SAM). The DM employs dilatation convolution with different dilation rates, which fully aggregates contextual feature information and reduces the loss of spatial resolution. The SAM takes the key information in the global features captured via the DM to focus and enhance the expression of the target features. The LSDAM receives the spatial-temporal features and sends them first to the DM to expand the sensory field, then subsequently to the SAM to extract the key information, and finally to the global, low-noise contextual information. Then, the enhanced features are dimensionally aligned with the spatially separated features and fed into the MSFFU for feature fusion. The MSFFU module refines the feature information via multiscale fusion and reduces the redundancy of the fused features, which enables the model to better understand the information in the whole scene. The MSFFU performs multiple levels of feature extraction for the double-branching features by utilizing different DO-Conv structures, and the individual MSFFU uses different DO-Conv structures to extract features from the dual-branch features at multiple levels, integrates each branch via element-by-element multiplication or addition, and then filters the irrelevant information in the features via a convolution operation. DO-Conv can accelerate the convergence of the network and improve the generalizability of the model, thereby improving the training speed of the model. Finally, the features at different levels are fed into the anchor-less frame-based prediction head for STAD. **Result** Comparative detection results between EAD and existing algorithms were analyzed for the public datasets UCF101-24 and atomic visual actions (AVA) version 2.2. For the UCF101-24 dataset, frame-mAP, video-mAP, frame per second(FPS), GFLOPs, and Params are used as evaluation metrics to assess the accuracy and spatial-temporal complexity of the model. For the AVA dataset, frame-mAP and GFLOPs are used as evaluation metrics. In this work, ablation experiments on the EAD model are performed on the UCF101-24 dataset. The frame-mAP is 79.52%, and the video-mAP is 49.29% after the addition of the MSFFU, which are improvements of 0.41% and 0.14%, respectively, from the baseline. The frame-mAP is 80.96%, and the video-mAP is 49.72% after the addition of the LSDAM, which are improvements of 1.85% and 0.57%, respectively, from the baseline. The final model of the EAD frame-mAP is 80.93%, the video-mAP is 50.41%, the FPS is 53 f/s, the number of GFLOPs is 2.77, and the number of parameters is 10.92 M, which are improvements of 1.82% and 1.26%, respectively, compared with the baseline frame-averaged accuracy and video-averaged accuracy. The leakage and misdetection phenomenon of the baseline method also improved. In addition, EAD is compared with existing real-time STAD algorithms, in which the frame-mAP is improved by 0.53% and 0.43%, and the GFLOPs are reduced by 40.93 and 0.13 compared with those of YOWO and YOWOv2, respectively. On the AVA dataset, the frame-mAP and GFLOPs reach 13.74% and 2.77, respectively, for an input frame count of 16; moreover, the frame-mAP and GFLOPs reach 15.92% and 4.4%, respectively, for an input frame count of 32. Compared with other mainstream methods, EAD uses a lighter backbone network to achieve lower computational costs while achieving impressive results. **Conclusion** This study proposes an STAD model called EAD, which is based on a two-stream network, to address the problems of missing global contextual information about actors' interactions and the poor characterization of fused features. The results of the experimental results of the proposed model on the UCF101-24 and AVA datasets verify its robustness and effectiveness in STAD tasks by comparing it with the baseline and current mainstream methods. The proposed model can also be applied to the fields of intelligent monitoring, automatic driving, intelligent medical care, and other fields.

Key words: deep learning; spatial-temporal action detection(STAD); two-branch network; dilated augmented module (DAM); multi-scale fusion

0 引 言

时空动作检测(spatial-temporal action detection, STAD)是视频理解领域中的重要任务之一,旨在对视频中发生的动作进行时间和空间的定位,同时对相关的动作类别进行分类。随着深度学习的快速发展,STAD受到越来越多研究者的重视,并在视频监控、视频字幕以及事件检测等方面得到了广泛应用(Venugopalan等,2015;Gan等,2015)。

由于视频处理的复杂性,早期的方法(Gkioxari和Malik,2015;Wang等,2016)将逐帧处理的检测结果整合成通道,以确定目标人物的动作类别及时空位置。然而,这些方法由于时间维度的分离而无法充分学习时间信息。一些研究者针对时间的连续性特点,利用光流图和3D卷积操作处理时间信息。Kalogeiton等人(2017)同时对连续K帧图像序列和光流图提取特征信息,减少动作预测的模糊性。Li等人(2020)将动作实例建模为每一帧动作中心点沿时序的运动轨迹,实现高效而精准的动作检测。Hou等人(2017)首次提出了将3D CNN(3D convolution neural network)应用到动作检测任务的算法T-CNN。Pan等人(2021)建立了一种包含高阶关系推理算子(high-order relation reasoning operator, HR_2O)和Actor-Context特征库的模型ACAR(actor-context-actor relation),有效地推理人物在复杂场景中的行为。王东祺和赵旭(2022)提出了类别敏感的全局时序关联动作检测模型,处理不同类别之间边界模式和动作模式差异大的问题。Zhao等人(2022)提出了一种将3D CNN与Transformer结合的端到端动作检测算法Tuber,以预测具有更灵活空间和时间范围的动作。贺楚景等人(2023)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,对行动者之间的交互作用和动作类别间的语义相关性进行建模。这些方法能够有效地捕获时空信息,实现较好的检测结果,但高质量光流图和3D卷积操作会给模型带来较大的计算开销。为此,一些研究者引入紧凑型神经网络处理时空动作检测任务。Köpüklü等人(2019)利用双流网络思想构建了单阶段的时空动作检测模型YOWO(you only watch once),并使用注意力机制和通道融合方法进行特征融合,以较快的检测速度实现高效的动作检测。Yang和Dai(2023)采用了现有的高效架构3D CNN来减轻计算开销,并构建了一个带有特征金字塔网络和解耦头的二维骨干,以提取不同层次的分类特征和定位特征,从而获得较高的检测精确度。

尽管上述方案取得了良好的效果,但仍然存在以下不足:1)现有方法大多依赖主干网络对视频片段进行特征建模,仅能捕捉到局部特征,忽略了与行动者交互的全局上下文信息,从而影响模型对整个场景信息的理解。场景中行动者的动作推理需要通过整合行动者的视觉特征、动作特征以及全局上下文信息来实现,其中全局特征和局部特征对模型推理都至关重要。2)基于双流网络的方法仅采用简单的逐通道连接处理双分支网络融合,其融合特征存在冗余,且各分支特征具有一定的语义差异,导致检测准确度较低。现有方法缺乏对检测速度和准确度的综合考虑,难以实现精度和速度的高效平衡。

为了解决上述问题,本文提出一种考虑全局上下文信息和特征融合的实时时空动作检测模型EAD(efficient action detector),建模整个场景中的关键特征。为弥补卷积神经网络特征建模的局部性,本文构建了轻量级空间扩张增强模块(lightweight spatial dilated augmented module, LSDAM)。LSDAM模块主要由扩张模块(dilated module, DM)和空间增强模块(spatial augmented module, SAM)两个子模块组成。其中,DM模块主要通过扩张卷积提升时空特征的全局性,SAM模块对全局特征进行加权,获取增强后的关键特征。为减少融合特征的冗余,本文构建了多尺度特征融合单元(multiscale feature fusion unit, MSFFU),利用不同的DO-Conv结构对双分支特征进行多个层次特征提取,并将各分支通过逐元素乘法或加法整合,随后通过卷积操作过滤特征中的无关信息。此外,本文采用紧凑型神经网络作为主干网络,并且MSFFU模块中引入的DO-Conv结构对EAD模型训练过程中的收敛速度有所提升,两者共同作用有效提升了整个模型的检测速度。总体而言,本文的贡献主要如下:1)提出一种实时时空动作检测模型EAD,将时空增强信息和空间信息进行高效整合,提升实时动作检测精度。2)针对全局上下文信息缺失问题,构建了LSDAM模块增大感受野,以学习全局性的强语义、高层次时空信息;此外提出了MSFFU模块,以较低的计算量实现双分支特征的高效融合。3)大量实验表明,在UCF101-24和

AVA(atomic visual actions)数据集上，EAD在检测速度和精度方面都表现出色。

1 方 法

在本节中，首先介绍EAD的整体架构并详细叙述推理流程。该模型的主要目标是实现轻量、高效的视频片段特征提取，在此基础上解决现有方法缺乏全局上下文交互信息以及双分支融合特征表达能力弱的问题，以提高动作检测的准确性。然后，对EAD模型中构建的各个子模块进行详细剖析，并介绍损失函数。

1.1 整体概述

为提升时空动作检测的检测精度和速度，本文提出了一种考虑全局上下文信息和特征融合的实时时空动作检测模型EAD，整体框架如图1所示。由于现有行为检测方法中基于2D CNN的方法需要依靠光流图实现高性能的行为检测，基于3D CNN的方法可以有效提取视频片段中的时空特征，但与光流图一样对计算资源要求较高，而本文方法采用双分支结构，使用轻量级的2D CNN和3D CNN作为主干建模视频片段的时空特征，同时实现了降低模型计算成本、提升模型检测速度的效果。EAD模型主要包括3个关键部分：2D分支、3D分支和融合头。其中，2D分支由经过预训练的2D主干网络、特征金字塔和解耦头组成，3D分支由3D主干网络和增强模块组成，融合头由MSFFU和预测头组成。EAD的推理流程如下：

1)将关键帧 $\boldsymbol{I}_K \in \mathbf{R}^{3\times H\times W}$ 输入经过预训练的2D分支网络对场景中的行动者进行检测，得到空间解耦特征，分别为分类特征 $\boldsymbol{F}_{\text{cls}}=\{\boldsymbol{F}_{\text{cls}_i}\mid \boldsymbol{F}_{\text{cls}_i}\in \mathbf{R}^{C_{\sigma 1}\times H/2^{i+2}\times W/2^{i+2}}\}_{i=1}^{3}$ 和定位特征 $\boldsymbol{F}_{\text{reg}}=\{\boldsymbol{F}_{\text{reg}_i}\mid \boldsymbol{F}_{\text{reg}_i}\in \mathbf{R}^{C_{\sigma 1}\times H/2^{i+2}\times W/2^{i+2}}\}_{i=1}^{3}$，其中，$W$、$H$为初始输入视频帧的宽高，$C_{\sigma 1}$为空间解耦特征的通道数，$i$为分支数。同时使用经过预训练的视频主干网络对视频片段 $\boldsymbol{V}=\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_K\}$ 提取视频时空特征 $\boldsymbol{F}_{\text{ST}}\in \mathbf{R}^{C_{\sigma 2}\times T\times H/32\times W/32}$，其中，$C_{\sigma 2}$、$T$分别表示 $\boldsymbol{F}_{\text{ST}}$ 的通道数和时间维度大小。

2)根据分组数g将时空特征 $\boldsymbol{F}_{\text{ST}}$ 逐通道划分为g个子特征，并分别送入LSDAM模块中进行特征增强处理，提取全局、低噪声的时空特征，再将分组特征整合得到 $\bar{\boldsymbol{F}}_{\text{ST}}\in \mathbf{R}^{C_{\sigma 2}\times T\times H/32\times W/32}$。

3)对 $\bar{\boldsymbol{F}}_{\text{ST}}$ 进行线性插值操作得到 $\bar{\boldsymbol{F}}$，使其在空间维度上与 $\boldsymbol{F}_{\text{cls}}$ 和 $\boldsymbol{F}_{\text{reg}}$ 对齐。随后将尺度对齐的时空

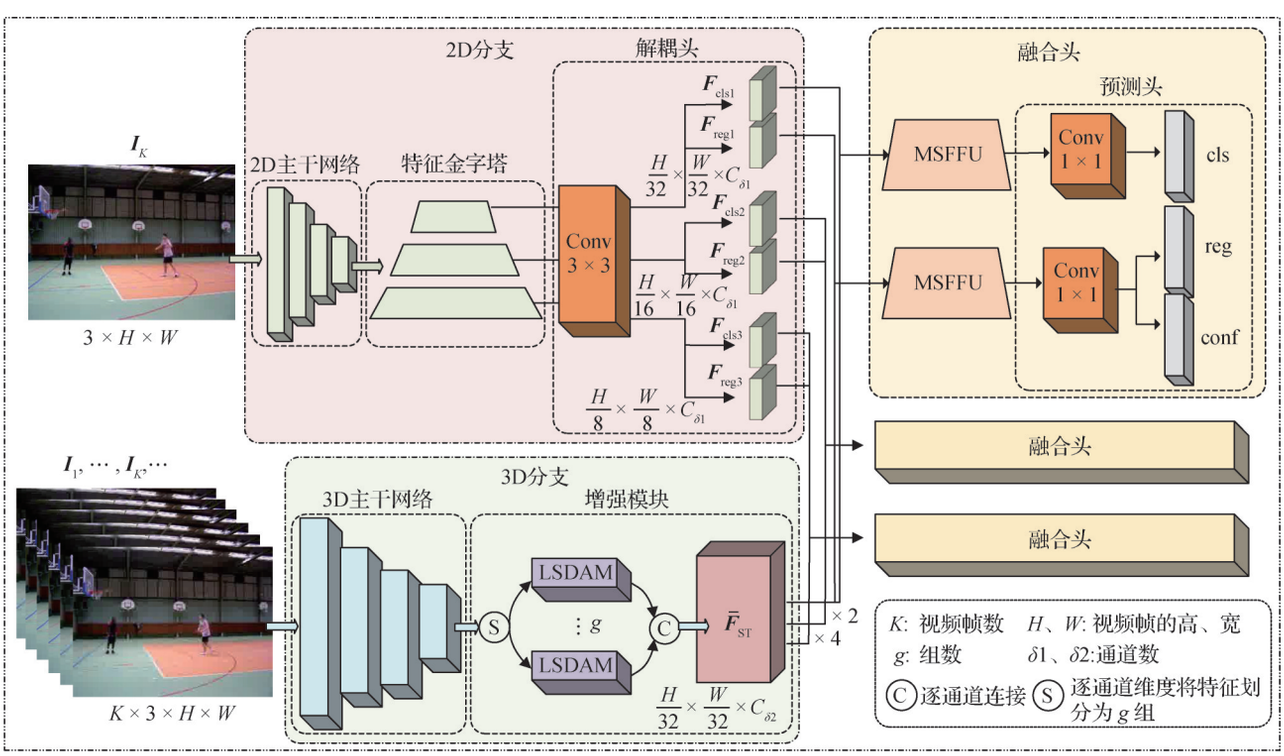


图1 EAD的整体框架

Fig. 1 Overall framework of EAD

特征与空间解耦特征分别送入MSFFU模块中，如图1所示。MSFFU模块利用条形卷积结构和不同的DO-Conv结构，提取不同尺度特征进行特征融合。然后经过卷积操作，过滤特征$\boldsymbol{F}$中的无关信息，增强特征表达能力。最后，送入基于无锚框的预测头，3个尺度的预测结果均作为预测结果，进而实现更加准确的动作类别和时空位置预测。

1.2 轻量空间扩张增强模块LSDAM

由于视频信息处理的复杂性，仅通过主干网络提取的特征中的全局交互信息并不完整，从而影响最终的检测精度。针对上述问题，本文设计LSDAM模块旨在捕获特征的全局相互依赖关系。如图2所示，LSDAM模块主要包括扩张模块（DM）和空间增强模块（SAM），DM采用不同扩张率的膨胀卷积（Wang等，2018），充分聚合上下文特征信息并降低空间分辨率损失；SAM将通过DM捕获到的全局特征中的关键信息进行聚焦，提升目标特征的表达能力。考虑到本文动作检测模型的实时性需求，模型采用分组思想以较少的参数量和计算量处理时空特征$\boldsymbol{F}_{\mathrm{ST}}$，LSDAM的推理流程如下：

将$\boldsymbol{F}_{\mathrm{ST}}$沿通道维度划分成$g$段特征，以一组特征$\boldsymbol{F}_{\tilde{n}} \in \mathbf{R}^{C_{\sigma 2}/g \times T \times H/32 \times W/32}$为一个LSDAM模块的输入。

1）引入膨胀卷积构建DM模块，以扩展特征的空间感受野，从而捕获全局的特征信息$\boldsymbol{F}'_{\tilde{n}} \in \mathbf{R}^{C_{\sigma 2}/g \times 4 \times T \times H/32 \times W/32}$。具体操作为

$$\boldsymbol{F}'_{\tilde{n}} = Concat(\boldsymbol{F}_{\tilde{n}}, Conv2D^{1}(\boldsymbol{F}_{\tilde{n}}), Conv2D^{2}(\boldsymbol{F}_{\tilde{n}}), Conv2D^{3}(\boldsymbol{F}_{\tilde{n}})) \tag{1}$$

式中，$Conv2D^{1}$表示内核大小为3×3、填充为1、组数为$C_{\sigma 2}/g$、膨胀系数为1的膨胀卷积。类似地，$Conv2D^{2}$和$Conv2D^{3}$的内核大小和组数与$Conv2D^{1}$相同，填充和膨胀系数取值相同，分别为2和3的膨胀卷积。$Concat$表示逐通道连接。

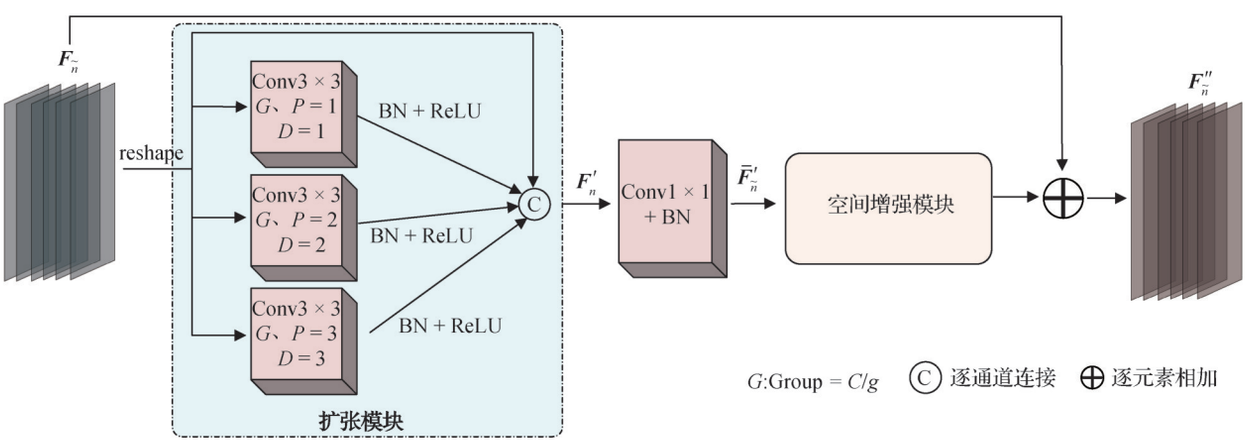


图2 轻量空间扩张增强模块

Fig. 2 Lightweight spatial dilated augmented module

2）$\boldsymbol{F}'_{\tilde{n}}$经过一个1×1的卷积运算降低特征冗余，实现跨通道信息的交互与整合，得到全局性特征图$\bar{\boldsymbol{F}}'_{\tilde{n}}$，具体为

$$\bar{\boldsymbol{F}}'_{\tilde{n}} = BN(Conv^{1\times 1}(\boldsymbol{F}'_{\tilde{n}})) \tag{2}$$

式中，$\bar{\boldsymbol{F}}'_{\tilde{n}} \in \mathbf{R}^{C_{\sigma 2}/g \times T \times H/32 \times W/32}$，$BN$为批归一化操作。

3）将$\bar{\boldsymbol{F}}'_{\tilde{n}}$送入SAM中，得到增强后的特征$\boldsymbol{F}''_{\tilde{n}}$。SAM处理流程如图3所示。

首先，通过卷积核大小为3×3的分组卷积运算，增强跨通道信息的交互；其次，经过3×3的最大池化层和卷积核大小为1×1的卷积运算，在通道维度上对特征向量进行通道压缩；随后，经过softmax操作，获得归一化后的权重$\boldsymbol{SA}_{\tilde{n}}$；最后，运用权重$\boldsymbol{SA}_{\tilde{n}}$对$\boldsymbol{F}_{\tilde{n}}$在空间维度上加权，获取增强特征$\boldsymbol{F}''_{\tilde{n}}$。

$$\boldsymbol{SA}_{\tilde{n}} = \mathrm{softmax}(Conv2D^{K=1}(MaxPool^{3\times 3,1}(Conv2D^{K=3}(\bar{\boldsymbol{F}}'_{\tilde{n}})))) \tag{3}$$

$$\boldsymbol{F}''_{\tilde{n}} = (\boldsymbol{SA}_{\tilde{n}} \otimes \boldsymbol{F}_{\tilde{n}}) \oplus \boldsymbol{F}_{\tilde{n}} \tag{4}$$

式中，$Conv2D^{K=3}$表示内核大小为3×3、填充为1、组数为$C_{\sigma 2}/g$的2D卷积层。$MaxPool^{3\times 3,1}$表示内核大小为3×3、填充为1的最大池化层。$Conv2D^{K=1}$表示内核大小为1×1、输出通道数为1的逐点卷积。⊗表示逐元素相乘，⊕表示逐元素相加。

4）通过Concat整合所有分组的特征，得到最终的时空特征$\bar{\boldsymbol{F}}_{\mathrm{ST}} \in \mathbf{R}^{C_{\sigma 2}\times T \times H/32 \times W/32}$。此外，每个卷积操作后都会应用BN（batch normalizaton）和ReLU（rectified linear unit），以防止模型仅执行线性操作，导致

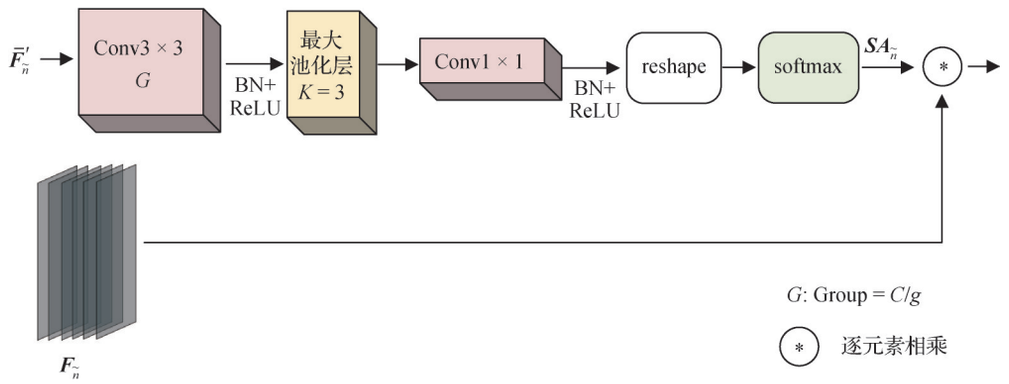


图3　空间增强模块

Fig. 3　Spatial augmented module

拟合能力不足。

1.3　多尺度特征融合单元 MSFFU

EAD 主要分为两条主分支，分别是负责输出空间解耦特征 $\boldsymbol{F}_{\mathrm{cls}}$ 和 $\boldsymbol{F}_{\mathrm{reg}}$ 的 2D 目标检测分支和输出时空特征 $\bar{\boldsymbol{F}}_{\mathrm{ST}}$ 的 3D 视频分支。为了保持两条分支特征在空间维度对齐，通过最近邻插值法对时空特征 $\bar{\boldsymbol{F}}_{\mathrm{ST}}$ 进行上采样操作，得到多尺度的时空特征 $\bar{\boldsymbol{F}} = \{\bar{\boldsymbol{F}}_i \mid \bar{\boldsymbol{F}}_i \in \mathbf{R}^{C_{\sigma 2} \times T \times H/2^{i+2} \times W/2^{i+2}}\}_{i=1}^{3}$。

现有方法通常采用通道连接或求和的方式实现特征融合，实现流程简单，但融合所得的信息存在冗余。因此，本文构建了 MSFFU 模块，旨在通过多尺度融合完善特征信息，帮助模型更好地理解整个场景中的信息。

MSFFU 模块主要由不同卷积核大小的 DO-Conv（Cao 等，2022）组成，DO-Conv 是由一个权重为 D 的逐深度卷积和一个权重为 W 的常规卷积组成，可以加快网络收敛速度并提高模型的泛化能力，具体卷积操作流程如图 4 所示。MSFFU 模块的具体流程如图 5 所示。

1）将空间解耦特征 $\boldsymbol{F}_{\mathrm{cls}}$ 和 $\boldsymbol{F}_{\mathrm{reg}}$ 与上采样后的时空增强特征 $\bar{\boldsymbol{F}}$ 分别沿通道维度进行连接得到 $\boldsymbol{F}$，并送入到 MSFFU 模块中。

2）MSFFU 模块分为两个阶段。

第 1 阶段将 $\boldsymbol{F}$ 经过 4 个不同的 DO-Conv 卷积结构，提取不同尺度特征，通过逐元素相乘和相加操作对不同尺度的特征进行整合。随后经过卷积操作，过滤特征 $\boldsymbol{F}$ 中的无关信息，增强特征表达能力。其中 DO-Conv 结构主要有 3 种：

第 1 种：由单个卷积核大小为 1 × 1 的 DO-Conv 卷积、GN（group normalization）和 GELU（Gaussian error linear units）（Cheng 等，2023）组成的 DO-Conv 结构，主要作用是实现跨通道的信息交互，具体为

$$\boldsymbol{F}_1 = GELU(GN(DOConv^{1\times 1}(\boldsymbol{F}))) \tag{5}$$

第 2 种：由两个卷积核大小分别为 1 × 3 和 3 × 1 的 DO-Conv 卷积、组归一化（GN）和激活函数 GELU 组成的条形卷积结构，用于过滤高层次特征的边界无关信息，具体为

$$\boldsymbol{F}_{2,3} = GELU(GN(DOConv^{3\times 1}(DOConv^{1\times 3}(\boldsymbol{F})))) \tag{6}$$

第 3 种：由 3 个卷积核大小为 3 × 3 的 DO-Conv 卷积、GN 和 GELU 组成的 DO-Conv 结构，以较少的参数提取更深层次的特征映射，具体为

$$\boldsymbol{F}_4 = GELU(GN(DOConv^{3\times 3}(DOConv^{3\times 3}(DOConv^{3\times 3}(\boldsymbol{F}))))) \tag{7}$$

在 DO-Conv 结构中引入组归一化 GN，其原因是它的计算与批量大小无关，且精度在很宽的批量大小范围内更加稳定。此外，与 ReLU 相比，GELU 可以避免梯度消失问题、加快模型收敛速度，为此在 DO-Conv 结构中引入了激活函数 GELU。

第 2 阶段将不同尺度特征进行融合补充。将经过第 1 阶段的卷积作用后的特征映射进行相乘和相加融合，再通过一个 1 × 1 的 DO-Conv 卷积，实现降维和跨通道信息交互整合。随后，通过跳跃连接将 $\boldsymbol{F}$ 与其相加，避免经过多层卷积运算后出现梯度消失现象，增强模型的泛化能力。最后，经过一个 1 × 1 卷积层得到最终的多尺度特征信息 $\boldsymbol{Out} = \{\boldsymbol{Out}_i \mid \boldsymbol{Out}_i \in \mathbf{R}^{dim \times H/2^{i+2} \times W/2^{i+2}}\}$，其中 dim 指进入分类头的通道尺寸。

1.4　损失函数

为了更好地拟合训练数据与优化模型，本文将

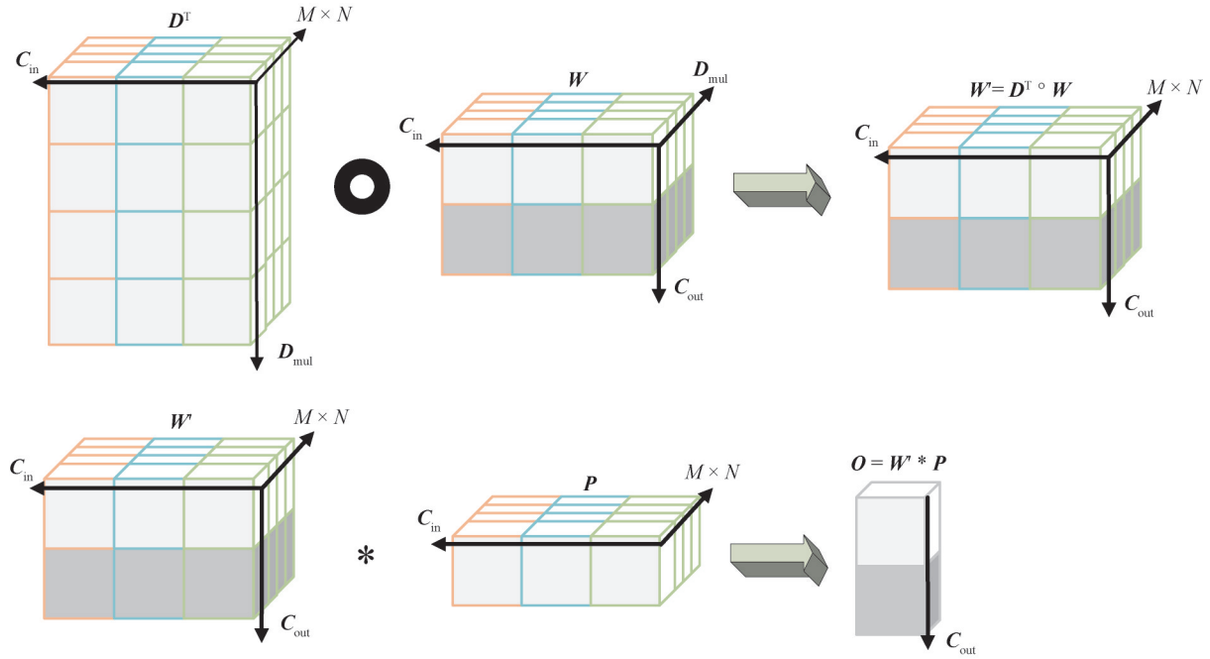


图4　DO-Conv 的具体操作流程

Fig. 4　Specific operations of depthwise over-parameterized convolution

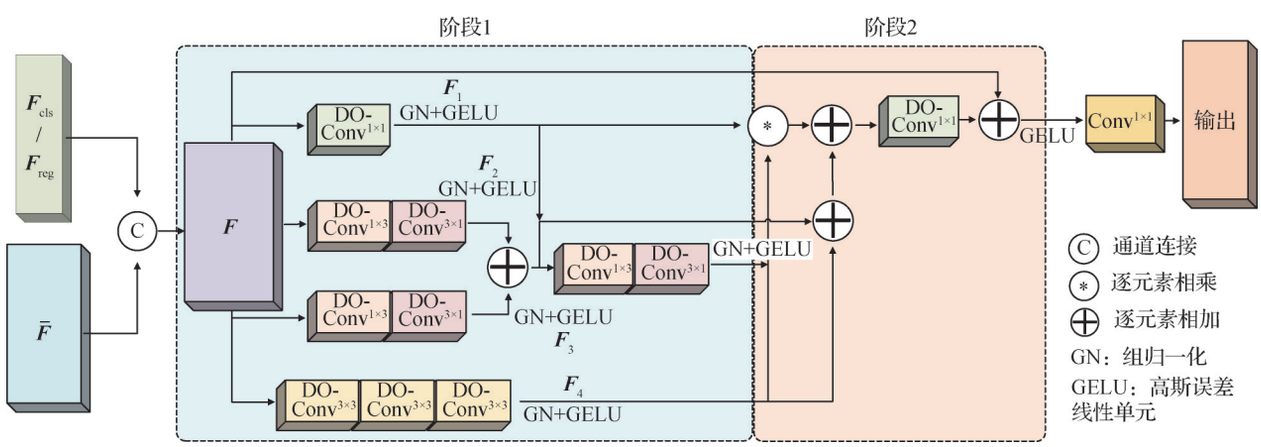


图5　多尺度特征融合单元

Fig. 5　Multiscale feature fusion unit

损失分为3个部分，分别为分类损失、定位损失和置信度损失。整体的损失定义为

$$\begin{aligned} L(\{cls_{x,y}\},\{reg_{x,y}\},\{conf_{x,y}\}) = \\ \frac{1}{N_{pe}}\sum_{x,y}L_{conf}(\bar{conf}_{x,y}, conf_{x,y}) + \\ \frac{1}{N_{pe}}\sum_{x,y}\alpha L_{cls}(\bar{cls}_{x,y}, cls_{x,y}) + \\ \frac{\lambda}{N_{pe}}\sum_{x,y}\alpha L_{reg}(\bar{reg}_{x,y}, reg_{x,y}) \end{aligned} \tag{8}$$

式中，$cls_{x,y}$、$reg_{x,y}$、$conf_{x,y}$分别为分类预测值、定位预测值和置信度预测值，$\bar{cls}_{x,y}$、$\bar{reg}_{x,y}$、$\bar{conf}_{x,y}$为真实值，N_{pe}为正样本数，λ为损失平衡因子，α在$\bar{cls}_{x,y} > 0$时取值为1，反之取值为0。

分类损失和置信度损失采用二进制交叉熵损失训练。定位损失采用SIoU损失(Gevorgyan, 2022)。此前的IoU(intersectio over union)损失函数未考虑真实框与预测框之间的方向，导致收敛速度较慢，对

此SIoU损失函数重新定义了相关损失函数，具体由以下4个部分组成：

1）角度损失

$$\Lambda = 1 - 2 \times \sin^2\left(\arcsin\left(\frac{c_{h1}}{\sigma}\right) - \frac{\pi}{4}\right) \tag{9}$$

式中，c_{h1}为真实框与预测框中心点的高度差，σ为真实框与预测框中心点的距离。

2）距离损失

$$\Delta = \sum_{t=x,y} 1 - \mathrm{e}^{-\gamma\rho_t} \tag{10}$$

$$\rho_x = \left(\frac{b_{c_x}^{\mathrm{gt}} - b_{c_x}}{c_{w2}}\right)^2,\ \rho_y = \left(\frac{b_{c_y}^{\mathrm{gt}} - b_{c_y}}{c_{h2}}\right)^2,\ \gamma = 2 - \Lambda$$

式中，$(b_{c_x}^{\mathrm{gt}}, b_{c_y}^{\mathrm{gt}})$为真实框的中心坐标，$(b_{c_x}, b_{c_y})$为预测框的中心坐标，$c_{w2}$，$c_{h2}$分别为真实框和预测框最小外接矩形的宽和高。

3）形状损失

$$\Omega = \sum_{t=w,h} (1 - \mathrm{e}^{-w_t})^{\theta} \tag{11}$$

$$w_w = \frac{|w - w^{\mathrm{gt}}|}{\max(w,\ w^{\mathrm{gt}})},\ w_h = \frac{|h - h^{\mathrm{gt}}|}{\max(h,\ h^{\mathrm{gt}})}$$

式中，w，h和w^{gt}，h^{gt}分别为预测框和真实框的宽和高，θ控制对形状损失的关注程度。

4）IoU损失

$$IoU = \frac{|\boldsymbol{B} \cap \boldsymbol{B}^{\mathrm{GT}}|}{|\boldsymbol{B} \cup \boldsymbol{B}^{\mathrm{GT}}|} \tag{12}$$

式中，$\boldsymbol{B}$和$\boldsymbol{B}^{\mathrm{GT}}$分别为预测框和真实框。

SIoU损失函数最终的完整定义为

$$L = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{13}$$

2 实　验

2.1 数据集和评价指标

本文主要在UCF101-24和AVA数据集上评估EAD模型。UCF101-24是UCF101（Soomro等，2012）的一个子集，其中包含24个动作类别，这些类别涵盖了各种动作，包括体育运动、舞蹈等，共计3 207个视频，并提供了相应的时空注释。

AVA（Gu等，2018）数据集包含80个原子视觉动作，由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。动作标签在这430个15 min的视频片段中进行了密集标注，在时间和空间上都是局部的。

对于UCF101-24数据集，本文采用了frame-mAP、video-mAP、帧速率（frame per second，FPS）、GFLOPs（giga floating-point operations per second）和Params作为评价指标，以评估模型的准确度和时空复杂度，在第1次划分上进行所有实验。此外，对于AVA数据集，以frame-mAP和GFLOPs作为评价指标，验证EAD的有效性和泛化性，仅介绍AVA数据集2.2版本的结果。

2.2 实现细节

1）主干网络。为了节省训练时间，本文采用YOLOv7（you only look once version 7）和解耦头框架作为目标检测主干，并使用ShuffleNetv2-2.0x（Ma等，2018）作为3D视频主干。采用紧凑型3D CNN进行特征提取，以减轻模型的计算开销。

2）训练细节。本文的模型基于深度学习框架PyTorch1.10.0和CUDA11.3部署，所有实验在NVIDIA Geforce GTX 4080（16 GB）上完成。模型采用AdamW优化器以加速训练，初始学习率设置为0.000 1，权重衰减为0.005，批量大小设置为8，并采用梯度累积策略（每16次迭代执行一次梯度更新），在前500次迭代中，采用线性预热策略。在UCF101-24数据集上训练7个epoch，并分别在第2、3、4、5和6轮次将学习率减半。在AVA数据集上训练9个epoch，并分别在3、4、5和6个epoch将学习率减半。

3）推理细节。本文模型的目标检测分支输入为关键视频帧，视频分支输入为16帧RGB帧，视频帧大小裁剪成224 × 224像素。在推理过程中，目标主干中分支数i取值为3，LSDAM中分组数g取值为8，分类头的通道尺寸dim为64，损失平衡因子λ设置为5，θ值设置为4，使用置信度评分大于0.5的检测框。

2.3 消融实验

本文在UCF101-24数据集上进行了详细的消融实验，以研究模型中不同成分对于整个模型的影响。文中的基线仅由主干网络（YOLOv7 + decoupled head，ShuffleNetv2）和单层的动作分类器组成。

2.3.1 整体结构的消融实验

本节分析EAD中各个组成部分对整体性能的影响，结果如表1所示。EAD采用ShuffleNetv2作为主干网络，以实现低计算量、低参数量的动作检测，主要是因为ShuffleNetv2本身属于轻量级网络。针对轻量级网络对于特征建模能力的不足，本文构建

了LSDAM模块，采用分组思想降低增强模块的计算量，同时通过增强结构完善了全局信息，提升了模型的检测准确度。与基线结果相比，GFLOPs和Params增加了0.03和0.51 M，frame-mAP和video-mAP提升了1.85%和0.57%。结果表明，加入LSDAM模块提升了模型的检测准确率，同时并未产生较多的计算参数，对整体模型的检测速度影响较小。此外，在实现双分支特征融合的基础上设计了MSFFU模块，主要是针对通用特征融合方法的缺陷，构建多尺度DO-Conv结构完善特征信息，提升模型的检测速度。加入MSFFU模块后，GFLOPs和Params分别降低了0.06和0.18 M，frame-mAP和video-mAP分别提升了0.41%和0.14%。结果表明，MSFFU模块相比基线的融合方式进一步节约了计算资源，同时检测准确率也有所提升。最终模型的frame-mAP和video-mAP分别提升了1.82%和1.26%，验证了本文模型提出的特征多尺度融合模块和轻量级时空特征增强模块均是有效的。

同时，本文对比了基线和EAD在UCF101-24数据集中各类别的frame-mAP，如图6所示。本文模型在Basketball、GlofSwing、IceDancing和RopeClimbing等动作类别上得到明显提升。原因是EAD通过增强特征的全局上下文信息，补全了空间上下文信息，进一步证实了EAD模型的有效性。

2.3.2 LSDAM模块的消融实验

为验证LSDAM模块设计的合理性与有效性，分别将SAM、时间增强模块（temporal augmented module，TAM）和DM组成不同的组合，进行消融实验，结果如表2所示（表2中S、T、D分别表示SAM模块、

表1 在UCF101-24数据集上探索EAD模型设计的消融研究

Table 1 Ablation study exploring EAD model design on the UCF101-24 dataset

方法	frame-mAP/%	video-mAP/%	GFLOPs	Params/M
基线	79.11	49.15	2.74	10.41
基线 + MSFFU	79.52(+0.41)	49.29(+0.14)	2.68	10.23
基线 + LSDAM	**80.96(+1.85)**	49.72(+0.57)	2.77	10.92
基线 + MSFFU + LSDAM	80.93(+1.82)	**50.41(+1.26)**	2.77	10.92

注：加粗字体表示frame-mAP和video-mAP的最优结果，括号内表示新增模块的性能提升。

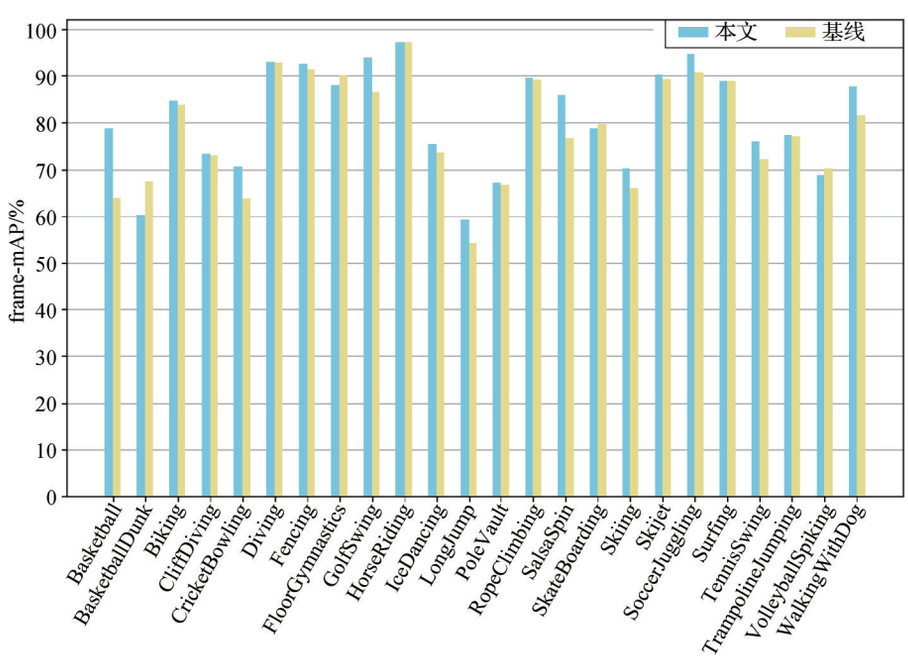


图6 在UCF101-24数据集上的frame-mAP

Fig. 6 Frame-mAP on the UCF101-24 dataset

TAM模块和DM模块)。

表2　在UCF101-24数据集上LSDAM的消融研究

Table 2　Ablation studies of LSDAM on the UCF101-24 dataset

/%

S	T	D	frame-mAP	video-mAP
–	–	–	79.11	49.15
√	–	–	80.11	49.04
–	√	–	80.95	49.70
√	√	–	80.03	**49.77**
√	√	√	80.66	49.60
√	–	√	**80.96**	49.72

注:加粗字体表示各列最优结果。“√”表示选用,“–”表示未选用。

SAM和TAM单独应用在时间和空间维度上进行特征建模,均能提升frame-mAP和video-mAP。而结合SAM和TAM时结果有所下降,主要原因是模型中的视频主干网络采用紧凑型卷积神经网络ShuffleNet,与ResNet、SlowFast网络相比,特征建模能力不足,经过多次的增强操作容易产生特征冗余,导致模型性能下降。将包含SAM、TAM和DM的结果与第4行结果比较,frame-mAP提升了0.63%。整合SAM与DM的整体表现最佳,frame-mAP和video-mAP分别达到了80.96%和49.72%。本文最终采用DM与SAM串联整合的方式构建了LSDAM模块。

2.3.3　MSFFU模块的消融实验

为验证MSFFU模块设计的有效性,对其进行不同融合方式的消融实验,结果如表3所示。本文中的基线采用Concat特征融合方法。与Concat相比,Sum在将待融合特征维度对齐过程中易丢失信息,致使模型性能下降。iAFF(iterative attentional feature fusion)(Dai等,2021)利用注意力机制为特征分配权重,允许融合具有不同语义和尺度的特征。而iAFF复杂的架构使其在模型中应用时,容易受到过拟合现象的影响,导致模型效果不佳。MSFFU模块在保持少量参数和计算量的同时将frame-mAP提升了0.41%,video-mAP提升了0.14%。

表3　在UCF101-24数据集上MSFFU的消融研究

Table 3　Ablation studies of MSFFU on the UCF101-24 dataset

/%

方法	frame-mAP	video-mAP
Concat	79.11	49.15
Sum	77.61	48.93
iAFF	78.90	49.27
MSFFU(本文)	**79.52**	**49.29**

注:加粗字体表示各列最优结果。

2.4　对比实验

2.4.1　UCF101-24数据集上的对比实验

本文将EAD分别与实时动作检测、时空动作检测中先进的算法在UCF101-24数据集上进行时空动作检测性能对比。在表4中,介绍了EAD与现有实时行为检测方法ACT(action tubelet detector)、MOC(movingcenter detector)、SAMOC(self-attention movingcenter detector)、YOWO(you only watch once)和YOWOv2(you only watch once v2)等在UCF101-24数据集上的对比结果。现有实时行为检测方法大多采用普通的卷积神经网络,如VGG(Visual Geometry

表4　在UCF101-24数据集上与其他实时行为检测器的比较

Table 4　Comparison with other real-time action detectors on the UCF101-24 dataset

方法	*K*	主干	frame-mAP/%	video-mAP/%	FPS/(帧/s)	GFLOPs	Params/M
ACT(Kalogeiton等,2017)	6	VGG	67.1	51.4	25	–	–
MOC(Li等,2020)	7	DLA-34	78.0	**53.8**	**53**	–	–
SAMOC(Ma等,2021)	5	DLA-34	79.3	52.5	40	–	–
YOWO(Köpüklü等,2019)	16	3D-ResNeXt-101	80.4	48.8	34	43.7	121.4
YOWOv2-T(Yang等,2023)	16	3D-ShuffleNetv2-1.0x	80.5	51.3	50	2.9	**10.9**
EAD(本文)	16	3D-ShuffleNetv2-2.0x	**80.93**	50.41	**53**	**2.77**	10.92

注:*K*表示批量大小,所有模型都在Kinetics-400上进行了预训练。加粗字体表示每列的最优结果,“–”表示原文未给出数据。

Group)、ResNet(residual networks)、ResNeXt(residual networks with next generation)等,它们的参数量都较大,从而导致训练速度减缓。EAD使用ShuffleNetv2作为主干网络,利用紧凑型神经网络低计算成本、低参数量的优点,实现更加高效的动作检测。EAD在检测性能和速度之间取得了更好的平衡,其FPS达到了53帧/s,frame-mAP和GFLOPs均达到了最优结果。与YOWOv2相比,EAD和YOWOv2采用相同的ShuffleNetv2主干,但本文模型检测速度和计算量更低。主要原因是YOWOv2利用逐通道连接并加入了通道注意力以实现双分支特征融合,逐通道连接后的特征维度较大,送入后续通道注意力操作处理时会产生较大的计算成本,进而影响模型检测速度。EAD针对普通特征融合方式的缺陷,构建了MSFFU模块,加快网络收敛速度并提高模型的泛化能力,同时减轻了LSDAM模块产生的参数对模型速度的影响,进一步提升了模型的计算速度,最终实现了更高的检测准确度,同时检测速度也得到提升。

表5展示了EAD与先进的时空检测算法在UCF101-24数据集上的对比结果。基于2D CNN方法需要结合光流提高性能,而光流获取需要离线操作,成本较高并降低检测速度。基于3D CNN方法取得良好性能的主要原因是骨干网络强大的特征建模能力。然而,基于双流网络的方法也可取得与基于3D CNN的方法相近的检测准确度,同时降低了模型复杂度。EAD以更少的计算量和参数实现了更好的性能,frame-mAP达到了80.93%,video-mAP达到了50.41%。

表5　在UCF101-24数据集上与最新技术的比较

Table 5　Comparison with state-of-the-art on the UCF101-24 dataset

/%

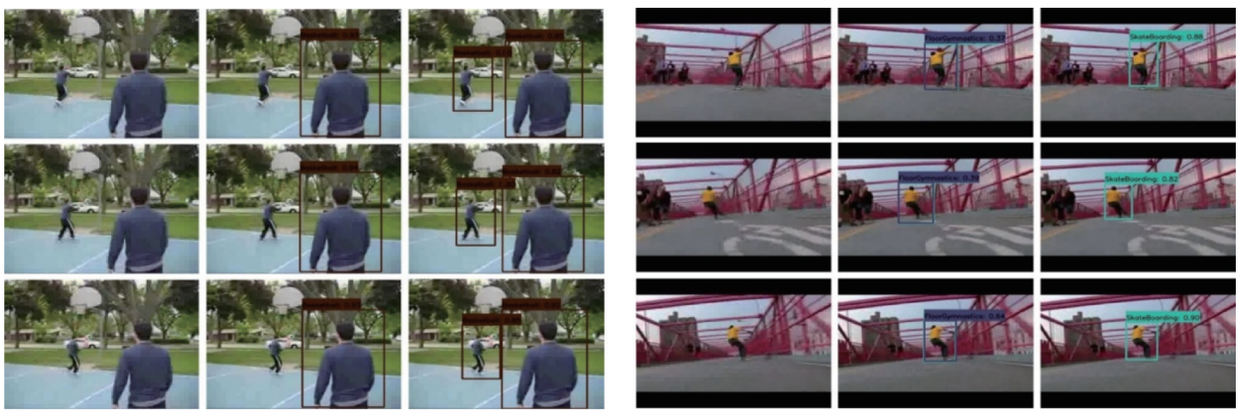
	方法	RGB视频帧	光流图像	frame-mAP	video-mAP
3D	T-CNN(Hou等,2017)	√	–	41.4	–
	I3D(Gu等,2018)	√	√	76.3	**59.9**
	ACAR(Pan等,2021)	√	–	82.4	–
	Tuber(Zhao等,2022)	√	–	83.2	58.4
	STMixer(Wu等,2023)	√	–	**83.7**	–
2D	ACT(Kalogeiton等,2017)	√	√	67.1	51.4
	TACNet(Song等,2019)	√	√	72.1	**54.4**
	MOC(Li等,2020)	√	–	73.1	51.0
	MOC(Li等,2020)	√	√	78.0	53.8
	SAMOC(Ma等,2021)	√	–	74.2	49.8
	SAMOC(Ma等,2021)	√	√	**79.3**	52.5
Two-stream	YOWO(Köpüklü等,2019)	√	–	80.4	48.8
	YOWOv2-T(Yang等,2023)	√	–	80.5	**51.3**
	EAD(本文)	√	–	**80.93**	50.41

注:加粗字体表示不同方法的最优结果。"–"表示原文未给出数据。

此外,本文展示了模型在UCF101-24测试集上的一些检测结果,如图7所示。每个子图中第1列为原始视频帧,第2列为基线模型的检测结果,第3列为本文模型的检测结果。在两个类别示例中,基线因仅依赖主干网络无法全面考虑场景中的完整信息,致使其检测结果存在误判和漏检现象。本文研究对动作的全局交互信息进行了增强,使其分类结果更准确,另外通过MSFFU融合两条分支的信息,帮助模型更好地理解整个场景中的信息,实现更准确的动作定位。

2.4.2　AVA数据集上的对比实验

AVA数据集是一个具有挑战性的数据集,其中场景是变化的,并且每个动作实例都标记了多个注释。目前,处理该数据集的大多数工作是基于3D



(a) 篮球　　(b) 滑板

图 7　UCF101-24 测试集的可视化结果

Fig. 7　Visualization results of the UCF101-24 test set((a)basketball;(b)skateborading)

CNN 的方法。表 6 展示了与基于 3D CNN 的方法，以及基于双流网络的方法在 AVA 数据集上的对比结果。基于 3D CNN 的方法表现出强大的时空动作检测性能，但其具有较高的时间复杂度，难以进行实时检测。与基于 3D CNN 的方法相比，基于双流网络的 EAD 方法在相对较低的计算成本下实现了更具竞争力的性能。EAD 计算成本低的主要原因是采用 ShuffleNetv2+YOLOv7 作为主干网络。由于 AVA 数据集中标签行为本身的复杂性，使 EAD 在与其他先进的动作检测器相比时，检测准确度上存在一定差距。与 YOWO、YOWOv2 相比，EAD 采用更轻量的主干网络，实现更低的计算成本。针对轻量主干的

表 6　在 AVA 数据集上与最新技术的比较

Table 6　Comparison with state-of-the-art on AVA dataset

方法	主干	输入	frame-mAP/%	GFLOPs
I3D(Gu 等，2018)	I3D-VGG	32 × 2	14.5	–
SlowFast(Feichtenhofer 等，2019)	SlowFast-R50	16 × 2	24.2	308
X3D-XL(Feichtenhofer，2020)	X3D-XL	16 × 5	26.1	290
WOO(Chen 等，2021)	SlowFast-R101	8 × 4	28.3	252
ACRN(Sun 等，2018)	S3D	8 × 8	17.4	–
ACAR(Pan 等，2021)	SlowFast-R101	8 × 8	**33.3**	–
Tuber(Zhao 等，2022)	I3D-ResNet50	16 × 4	26.1	132
Tuber(Zhao 等，2022)	I3D-ResNet101	16 × 4	28.6	246
Tuber*(Zhao 等，2022)	CSN-152	32 × 2	31.7	120
SE-STAD(Sui 等，2023)	SlowFast-R101	8 × 4	29.3	165
STMixer(Wu 等，2023)	SlowFast-R101	32 × 2	30.1	–
YOWO(Köpüklü 等，2019)	3D-ResNeXt-101 + Darknet-19	16 × 4	17.9	44
YOWO(Köpüklü 等，2019)	3D-ResNeXt-101 + Darknet-19	32 × 4	19.1	82
YOWOv2-T(Yang 和 Dai，2023)	3D-ShuffleNetv2-1.0x + YOLOv7	16 × 4	14.9	2.9
YOWOv2-T(Yang 和 Dai，2023)	3D-ShuffleNetv2-1.0x + YOLOv7	32 × 4	15.6	4.5
EAD(本文)	3D-ShuffleNetv2-2.0x + YOLOv7	16 × 4	13.74	**2.77**
EAD(本文)	3D-ShuffleNetv2-2.0x + YOLOv7	32 × 4	15.92	4.4

注：frame-mAP 表示 IoU 为 0.5 的帧平均准确度，所有模型都在 Kinetics-400 上进行了预训练。加粗字体表示各列最优结果。“–”表示原文未给出数据。

特征捕获能力不足的问题，构建LSDAM模块，在保持更低计算开销的同时取得了可观的结果，实现实时行为检测。

图8展示了AVA测试集中部分视频中关键帧的检测结果。第1列为测试集中两个片段的原始视频帧，第2列为测试集中两个片段的视频帧示例及真实目标标注，第3列为本文模型对于视频帧示例的动作检测结果。本文模型准确定位出视频片段中所有行动者，同时检测出行动者发出的简单动作。但对于复杂交互动作的判断存在一定误差，对于视频片段中时间交互的建模存在不足，未来将对其进一步研究。

图8　AVA测试集的可视化结果

Fig. 8　Visualization results of the AVA test set

3　结　论

针对行动者交互的全局上下文信息缺失、融合特征的表征能力差的问题，本文提出了一种基于双流网络的时空动作检测模型EAD。构建的LSDAM模块通过扩张结构增大感受野，提取全局信息，同时引入加强模块进行关键信息增强，有效地抑制无关信息对预测结果的影响，且采用分组思想处理增强结构，以产生较少的计算量与参数量，更好地平衡模型的检测准确度和检测速度。此外，构建的MSFFU模块通过融合多层不同的DO-Conv卷积结构，以提取多尺度目标特征，使模型更加全面地理解场景信息，进一步提高模型的检测准确度和收敛速度。本文在UCF101-24和AVA数据集上对模型进行实验，验证了其在时空动作检测任务中的鲁棒性和有效性。本文提出的EAD模型在一定程度上处理了现有方法中全局场景信息缺失的问题，但仍存在不足。未来的研究将针对如何提升建模复杂场景中交互关系的能力，进一步改善模型应对复杂场景的性能。

参考文献（References）

Cao J M, Li Y Y, Sun M C, Chen Y, Lischinski D, Cohen-Or D, Chen B Q and Tu C H. 2022. DO-Conv: depthwise over-parameterized convolutional layer. IEEE Transactions on Image Processing, 31: 3726-3736 [DOI: 10.1109/TIP.2022.3175432]

Chen S F, Sun P Z, Xie E Z, Ge C J, Wu J N, Ma L, Shen J J and Luo P. 2021. Watch only once: an end-to-end video action detection framework//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 8158-8167 [DOI: 10.1109/ICCV48922.2021.00807]

Cheng Y, Wang W, Ren Z P, Zhao Y F, Liao Y L, Ge Y, Wang J, He J X, Gu Y K, Wang Y X, Zhang W J and Zhang C. 2023. Multi-scale feature fusion and Transformer network for urban green space segmentation from high-resolution remote sensing images. International Journal of Applied Earth Observation and Geoinformation, 124: #103514 [DOI: 10.1016/j.jag.2023.103514]

Dai Y M, Gieseke F, Oehmcke S, Wu Y Q and Barnard K. 2021. Attentional feature fusion//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3559-3568 [DOI: 10.1109/WACV48630.2021.00360]

Feichtenhofer C. 2020. X3D: expanding architectures for efficient video recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 200-210 [DOI: 10.1109/CVPR42600.2020.00028]

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]

Gan C, Wang N Y, Yang Y, Yeung D Y and Hauptmann A G. 2015. DevNet: a deep event network for multimedia event detection and evidence recounting//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE:

2568-2577［DOI：10.1109/CVPR.2015.7298872］

Gevorgyan Z. 2022. SIoU loss：more powerful learning for bounding box regression.［EB/OL］.［2024-02-22］. https://arxiv.org/pdf/2205.12740.pdf

Gkioxari G and Malik J. 2015. Finding action tubes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston，USA：IEEE：759-768［DOI：10.1109/CVPR.2015.7298676］

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 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(5)：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(5)：1499-1512）［DOI：10.11834/jig.211040］

Hou R，Chen C and Shah M. 2017. Tube convolutional neural network（T-CNN）for action detection in videos//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice，Italy：IEEE：5823-5832［DOI：10.1109/ICCV.2017.620］

Kalogeiton V，Weinzaepfel P，Ferrari V and Schmid C. 2017. Action tubelet detector for spatio-temporal action localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice，Italy：IEEE：4415-4423［DOI：10.1109/ICCV.2017.472］

Köpüklü O，Wei X Y and Rigoll G. 2019. You only watch once：a unified CNN architecture for real-time spatiotemporal action localization［EB/OL］.［2024-02-22］. https://arxiv.org/pdf/1911.06644.pdf

Li Y X，Wang Z X，Wang L M and Wu G S. 2020. Actions as moving points//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：68-84［DOI：10.1007/978-3-030-58517-4_5］

Ma N N，Zhang X Y，Zheng H T and Sun J. 2018. ShuffleNet V2：practical guidelines for efficient CNN architecture design//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：122-138［DOI：10.1007/978-3-030-01264-9_8］

Ma X R，Luo Z G，Zhang X，Liao Q，Shen X Y and Wang M Z. 2021. Spatio-temporal action detector with self-attention//Proceedings of 2021 International Joint Conference on Neural Networks. Shenzhen，China：IEEE：1-8［DOI：10.1109/IJCNN52387.2021.9533300］

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］

Song L，Zhang S W，Yu G and Sun H B. 2019. TACNet：transition-aware context network for spatio-temporal action detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：11979-11987［DOI：10.1109/CVPR.2019.01226］

Soomro K，Zamir A R and Shah M. 2012. UCF101：a dataset of 101 human actions classes from videos in the wild［EB/OL］.［2024-02-22］. https://arxiv.org/pdf/212.0402.pdf

Sui L，Zhang C L，Gu L X and Han F. 2023. A simple and efficient pipeline to build an end-to-end spatial-temporal action detector//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa，USA：IEEE：5988-5997［DOI：10.1109/WACV56688.2023.00594］

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］

Venugopalan S，Rohrbach M，Donahue J，Mooney R，Darrell T and Saenko K. 2015. Sequence to sequence-video to text//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago，Chile：IEEE：4534-4542［DOI：10.1109/ICCV.2015.515］

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 L M，Qiao Y，Tang X O and Van Gool L. 2016. Actionness estimation using hybrid fully convolutional networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE：2708-2717［DOI：10.1109/CVPR.2016.296］

Wang P Q，Chen P F，Yuan Y，Liu D，Huang Z H，Hou X D and Cottrell G. 2018. Understanding convolution for semantic segmentation//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe，USA：IEEE：1451-1460［DOI：10.1109/WACV.2018.00163］

Wu T，Cao M Q，Gao Z T，Wu G S and Wang L M. 2023. STMixer：a one-stage sparse action detector//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：14720-14729［DOI：10.1109/CVPR52729.2023.01414］

Yang J H and Dai K. 2023. YOWOv2：a stronger yet efficient multi-level detection framework for real-time spatio-temporal action detection［EB/OL］.［2024-02-22］. http://arxiv.org/pdf/2302.06848.pdf

Zhao J J，Zhang Y Y，Li X Y，Chen H，Shuai B，Xu M Z，Liu C H，Kundu K，Xiong Y J，Modolo D，Marsic I，Snoek C G M and Tighe J. 2022. TubeR：tubelet Transformer for video action detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：13588-

13597［DOI: 10.1109/CVPR52688.2022.01323］

作者简介

程勇，男，教授，主要研究方向为计算机网络、软件工程与深度学习。E-mail：yongcheng@nuist.edu.cn
高园元，女，硕士研究生，主要研究方向为计算机视觉与动作识别检测。E-mail：202212490406@nuist.edu.cn
王军，男，教授，主要研究方向为软件工程、物联网、信息系统和物理信息融合。E-mail：wangjun@nuist.edu.cn
杨玲，女，副教授，主要研究方向为全程图像处理、信号与系统。E-mail：002303@nuist.edu.cn
许小龙，男，教授，主要研究方向为知识图谱、大数据与人工智能应用、边缘计算。E-mail：xlxu@ieee.org
程遥，男，硕士研究生，主要研究方向为计算机视觉与行为识别。E-mail：202212490425@nuist.edu.cn
张开华，男，教授，主要研究方向为计算机视觉与模式识别。E-mail：002530@nuist.edu.cn