引用格式：夏子瀛, 头旦才让, 张艺杰, 等. IIG-VSAD：基于实例信息引导的视频流行为检测方法 [J]. 电子科技大学学报, 2026, 55(1): 129-136.
XIA Z Y, THUPTEN T, ZHANG Y J, et al. IIG-VSAD: Instance information-guided video stream action detection[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(1): 129-136.

IIG-VSAD：基于实例信息引导的视频流行为检测方法

夏子瀛[1]，头旦才让[1,2]，张艺杰[1]，刘思宇[1]，邓昌健[1]，程　建[1*]，尼玛扎西[3]

（1. 电子科技大学信息与通信工程学院，成都 611731；2. 青海师范大学藏语智能全国重点实验室，西宁 810008；
3. 西藏大学信息科学技术学院，拉萨 850000）

摘要：视频流时序行为检测要求在仅观测到历史及当前时空信息的条件下，于在线视频流中准确检测当前时刻的行为类别。现有方法主要通过设计网络并利用帧级信息进行监督学习，对单帧信息过度敏感，缺乏时序一致性，导致检测准确性不足。针对以上问题，提出实例信息引导的视频流行为检测方法，在单帧检测基础上扩增实例信息，提出实例图推理策略生成导引，融合时序特征提升检测性能。于公开视频数据验证实验结果，证明了该方法的有效性且具备高效的检测效率。

关键词：视频流；行为检测；实例导引生成；实例图推理；注意力机制
中图分类号：TP391.4　　**文献标志码：**A　　DOI: 10.12178/1001-0548.2024328

IIG-VSAD: Instance information-guided video stream action detection

XIA Ziying[1], THUPTEN Tsering[1,2], ZHANG Yijie[1], LIU Siyu[1],
DENG Changjian[1], CHENG Jian[1*], and NYIMA Tashi[3]

(1. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. The State Key Laboratory of Tibetan Intelligence, Qinghai Normal University, Xining 810008, China;
3. School of Information Science and Technology, Xizang University, Lhasa 850000, China)

Abstract: Action detection in video streams requires accurately identifying the action category at the current moment within an online video stream, given only the historical and current spatiotemporal information observed up to that point. The existing methods mainly conduct supervised learning by designing networks and using frame-level information, which are overly sensitive to single-frame information and lack temporal consistency, resulting in insufficient detection accuracy. To address the aforementioned issues, an instance-guided video stream action detection method is proposed. Building upon frame-level detection, instance information is augmented, an instance graph reasoning strategy is proposed for generating guidance, and temporal features are then integrated to enhance detection performance. Finally, the proposed algorithm is validated on publicly available video datasets, and experimental results demonstrate the effectiveness of the method and its high detection efficiency.

Key words: video stream; action detection; instance guidance generation; instance graph reasoning; attention mechanism.

视频流行为检测要求在仅观测到历史及当前时空信息的条件下，对当前视频帧中的目标行为进行检测。视频流行为检测任务具有检测效率高、实时性强等特性，在人类日常生产生活中具有极高的应用价值，对安防监控[1]、自动驾驶[2]、具身智能[3]以及智能机械制造[4]等领域的技术发展具有极大推动意义。

对视频中的行为进行精准检测是视频理解领域

收稿日期：2024−12−02
基金项目：国家自然科学基金民航联合基金重点项目（U2233209）；国家自然科学基金青年项目（62306158）；四川省自然科学基金（2023NSFSC0484）
作者简介：夏子瀛，博士生，主要从事视频理解方面的研究。
*通信作者 E-mail: chengjian@uestc.edu.cn

的基础任务，近年来研究人员对相关理论进行了大量研究，其研究重点聚焦于如何对离线视频中的人类行为进行检测。与图像域中目标检测任务类似，离线行为检测框架可分为一阶段（One-Stage）、两阶段（Two-Stage）以及无锚框（Anchor-Free）3 类。如文献 [5] 中提出 ReAct 网络，在 One-Stage 的 DETR[6] 框架基础上设计关系注意力机制改进解码器实现行为检测。ContextLoc++[7] 提出 3 种子网络，同时增强 Two-Stage 框架下的局部、全局与多尺度特征。BMN[8] 通过生成边界匹配置信度图，依据图中时序响应实现无锚框行为检测。更近一步地，ADI-Diff[9] 将扩散过程引入行为检测任务，构建二值图扩散流程，实现准确且高效的行为检测。AdaTAD[10] 通过设计时序信息适配器降低模型参数，减少了端到端的训练负担。

与需要完整时空上下文信息的离线行为检测任务相比，视频流行为检测缺乏对未来信息的感知手段，同时需要对行为边界进行更加细粒度的检测，因此更具挑战性。如何有效利用有限的时序信息，抑制行为背景信息对检测模型的干扰，是实现精准检测的关键。现有方法在深度学习理论下，主要基于循环神经网络（RNN）以及 Transformer[11] 架构设计视频流行为检测器。前者主要依托 RNN 的长时记忆功能，自主保留历史视频帧中的行为信息作为辅助，结合当前视频帧信息实现有效行为检测。如文献 [12] 提出的 TRN 设计基于 RNN 的时序循环网络单元，预测未来行为并与当前行为信息融合，有效地实现了视频流行为检测。文献 [13] 分析了 RNN 结构缺陷的潜在原因，改进优化流程，提升了检测性能。由于 RNN 网络具有非并行计算以及梯度消失等缺陷，导致其网络难以优化。针对以上问题，近年来在线视频流行为检测方法开始基于 Transformer 架构进行设计，其中注意力机制能够直接实现全局信息获取，而非 RNN 结构的步进信息记忆，该特性有利于实现有效的在线检测性能。如文献 [14] 提出的 OadTR 网络，将单帧图像信息作为 Token 输入至编解码器，实现了检测性能与检测效率的同步提升。文献 [15] 则在 Transformer 架构基础上从知识蒸馏的角度解决视频流行为检测问题。

然而，现有方法聚焦于与任务目标对齐的单帧检测，导致模型对单帧信息过于敏感，进而忽视当前帧所在行为实例的上下文信息对检测的影响，导致算法模型检测性能下降。为改善上述问题，提出一种基于实例信息引导的视频流行为检测方法（instance information guided video stream action detection, IIG-VSAD）。该方法在以上基础上，显式地引入行为实例信息，从行为实例的角度显式强化已有帧级信息，从而提高检测结果时序一致性。其具体由时序特征编码器、时序导引生成器以及特征融合模块组成。首先基于注意力机制设计时序编码器，自适应聚合与当前帧相关的时序信息，提取时序特征。其次，设计实例引导分支，构建实例导引生成器，在帧级标注信息的基础上，构建实例预测目标函数，有效预测当前时刻所在的行为边界，并根据预测结果生成多个行为实例提名。此后，根据所预测的实例提名信息，提出实例关系图构建方法将实例提名集合映射为图，并通过设计图推理策略，自适应调整信息成行为掩码序列。最终设计特征融合模块，依赖掩码序列显式引导聚合时序特征，提取实例特征，最终实现对在线视频流行为的有效检测。

1 问题描述

为提升辨别性，首先对离线视频时序行为检测问题的定义进行阐述。给定训练视频集$\mathcal{V}_{\text{train}}=\{(\boldsymbol{V}_i,\boldsymbol{Y}_i)\}_{i=1}^{N}$，其中$V_i=\{x_1,x_2,\cdots,x_n\}$表示第$i$个视频数据，具有$n$帧，$N$表示视频集中的视频数量。而$I_i=\{(c_i,a_i)\}_{i=1}^{M}$表示第$i$个视频的行为实例标注，其中$c_i=\{1,2,\cdots,C\}$表示行为类别，共$C$个行为类别，$a_i$表示该实例在视频中的时间区域，$M$表示在该视频中出现的实例数量。在此前提下，离线行为检测要求对测试视频集$\mathcal{V}_{\text{test}}$中的行为实例进行精准检测。

与离线行为检测任务不同，视频流时序行为检测更具挑战性，要求给定一个在线视频流，基于过去到当前时间的观察信息，识别当前视频帧中的行为。具体为给定视频流$S=\{\cdots,x_{-1},x_0\}$，该流仅包含当前时刻以及历史的视频帧信息，不包含任何未来信息，要求准确预测当前时刻的行为y_t。在实际训练测试时，为减少时序上的信息冗余和计算量，首先将视频流分段得到视频片段，其中每个片段$s_i=\{x_{i-l+1},\cdots,x_i\}$，长度为$l$，以片段序列作为最小输入单元进行行为检测。

2 方法论

2.1 方法概述

本文提出的方法框架如图 1 所示，首先构建特

征提取器，对输入的片段序列进行特征提取，为了得到有效的特征表达，首先对每个片段s_i计算光流图。与文献 [14] 类似，随后使用与行为识别任务中预训练的 TSN[16] 网络进行片段特征提取，最终分别得到时序片段的视觉特征$\boldsymbol{f}_i^r$与光流特征$\boldsymbol{f}_i^f$，维度为 2 048。其次设计时序编码器提取时序特征；设计实例导引生成器提取生成实例提名并进行图推理生成时序掩码序列；设计特征融合模块，基于掩码序列信息进行引导并完成信息聚合得到实例特征；最终构建分类器实现视频流行为识别。下文将分别对所提出的时序特征编码器、实例导引生成器以及特征融合模块进行说明。

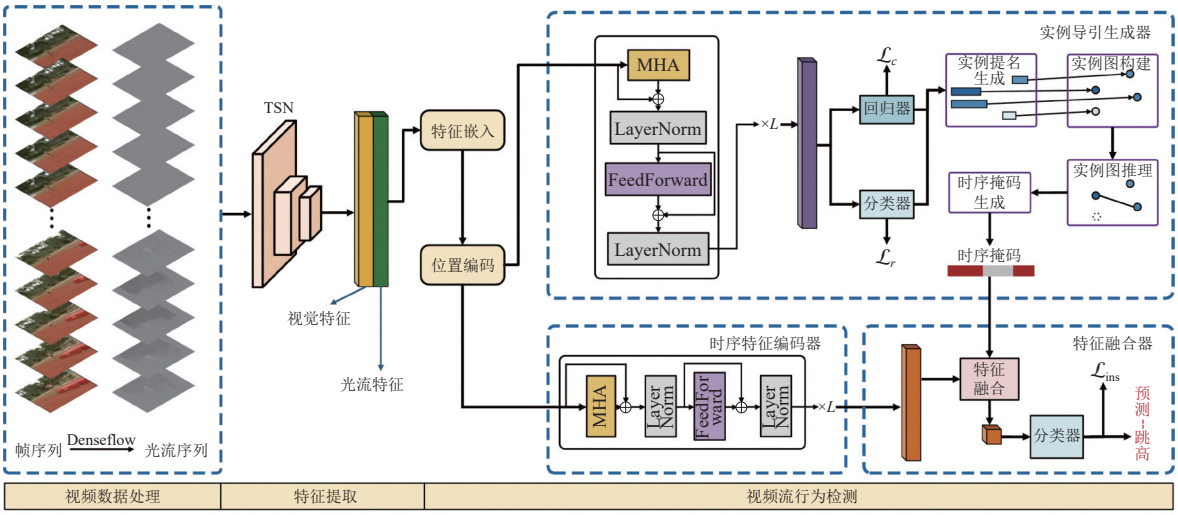

图 1　IIG-VSAD 方法框架

2.2　时序特征编码器

在线视频流行为检测任务要求保证实际推理的实时性，模型计算量会受到约束，将完整历史片段序列作为输入不现实。因此，期望在保证实时性的前提下，获取具备辨别性的时序特征。对此，提出基于注意力机制的时序编码器，结构如图 1 所示。输入长度为 T 的时序片段序列，对于每个片段，首先将视觉特征与光流特征在通道维度进行级联得到时空多模态特征$\boldsymbol{f}_i$。其次设计由线性层-LeakyReLU-Dropout 层级联得到的映射层emb(·)对时空多模态特征序列完成特征映射。

$$\boldsymbol{F}=\{\mathrm{emb}(\boldsymbol{f}_{-T+1}),\cdots,\mathrm{emb}(\boldsymbol{f}_0)\}\in\mathbb{R}^{T\times D}$$

式中，D 表示映射后的特征维度。

完成特征适应后，基于自注意力机制实现时序特征提取。为了保留各个时间点的时序因果信息，首先对特征序列$\boldsymbol{F}$进行时序位置编码，得到：

$$\boldsymbol{F}'=\boldsymbol{F}+\boldsymbol{F}_{\mathrm{pos}}$$

式中，$\boldsymbol{F}_{\mathrm{pos}}\in\mathbb{R}^{T\times D}$为可学习的位置向量。然后构建时序编码器$\mathrm{Enc}^F(\cdot)$，编码$\boldsymbol{F}'$得到时序特征$\boldsymbol{F}^{\mathrm{frame}}\in\mathbb{R}^{T\times D}$，其计算流程为：

$$\boldsymbol{F}^{\mathrm{frame}}=\mathrm{Enc}(\boldsymbol{F}')=\mathrm{norm}(\mathrm{Forward}(\boldsymbol{F}_A)+\boldsymbol{F}_A)$$
$$\boldsymbol{F}_A=\mathrm{norm}(\mathrm{MHA}(\boldsymbol{F}')+\boldsymbol{F}')$$

式中，norm(·)表示层归一化；Forward(·)表示由线性层-GELU 层-线性层级联而成的前馈网络。为了有效聚合与当前时刻相关的时序信息，采用多头注意力模块MHA(·)[11] 自主预测注意力图进行时序信息聚合。时序特征$\boldsymbol{F}^{\mathrm{frame}}$后续将与由实例导引生成器获取的时序掩码序列结合获取实例级特征。

2.3　实例导引生成器

实例导引生成器用于生成带有实例信息的时序掩码作为后续融合特征的导引。生成器首先基于时序特征与实例级优化策略生成大量实例提名，其次构建实例图并进行图推理实现对实例信息的高效利用，最终基于图信息生成时序掩码。

实例提名生成：在前面所得的时序特征$\boldsymbol{F}^{\mathrm{frame}}$基础上构建分类器预测行为类别，即能够初步实现在线视频流的分类。然而，在上述技术流程中，模型会过度聚焦于当前帧的信息，预测结果缺乏对当前视频帧所在行为实例的完整性考量，导致在整个视频流检测过程中结果缺乏一致性。为解决上述问题，提出一种实例导引生成器，如图 1 所示。首先构建与时序编码器结构相同的导引特征编码器

$\mathrm{Enc}^I(\cdot)$，对特征$\boldsymbol{F}'$进行特征编码，得到时序导引特征$\boldsymbol{F}^{\mathrm{prop}} \in \mathbb{R}^{T\times D}$。基于自注意力机制的导引特征编码器能够使输入特征获取全局上下文信息。之后，分别构建帧级分类器$\mathrm{Cls}^P(\cdot)$与帧级回归器$\mathrm{Reg}^P(\cdot)$，输入时序导引特征$\boldsymbol{F}^{\mathrm{prop}}$以获取行为预测概率序列$\boldsymbol{P}=\{\hat{\boldsymbol{p}}_{-T+1},\cdots,\hat{\boldsymbol{p}}_0\}\in\mathbb{R}^{T\times(C+1)}$与实例边界序列$\boldsymbol{B}=\{\hat{\boldsymbol{b}}_{-T+1},\cdots,\hat{\boldsymbol{b}}_0\}\in\mathbb{R}^{T\times 2}$。在预测概率中，第$C+1$类表示背景类，在边界序列中，$\hat{\boldsymbol{b}}_i=(\hat{s}_i,\hat{e}_i)$，其中$\hat{s}_i,\hat{e}_i$分别表示当前片段点$t_i$所在行为实例的起始时间与结束时间对于该时间点的相对偏移。在该生成器中，基于交叉熵损失作为分类损失$\mathcal{L}_c$，使用 DIoU [17] 损失$\mathcal{L}_{\mathrm{DIoU}}$作为回归损失$\mathcal{L}_r$进行优化，具体为：

$$\mathcal{L}_c=-\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{C+1}\sum_{i=1}^{C+1}y_i\log(\hat{p}_{t,i})\right)$$

$$\mathcal{L}_r=-\frac{1}{\sum_{i=1}^{T}\mathbb{I}(t_i\in I)}\sum_{i=1}^{T}\mathbb{I}(t_i\in I)\mathcal{L}_{\mathrm{DIoU}}(\hat{\boldsymbol{b}}_i,\boldsymbol{b}_i)$$

式中，$\boldsymbol{b}_i$表示第i个片段的实例边界偏移，若该片段不属于任何一个实例，则$\boldsymbol{b}_i=(-1,-1)$；$\mathbb{I}(\cdot)$表示指示函数。最终对于长度为T的片段序列，可获得元素数量为T的实例提名集$\hat{I}=\{(\hat{p}_i,\hat{s}_i,\hat{e}_i)\}_{i=1}^{\mathrm{T}}$。

实例图推理：不同于对片段序列做逐帧分类预测，基于回归生成的实例提名实现了片段在实例中的整体性考量，能够显式地聚合与当前时刻相关的信息。然而，生成的大量提名中不可避免地存在低质量提名，其中的噪声信息不利于信息聚合，无法直接利用。而如离线行为检测方法中简单地直接使用 Soft-NMS[18] 等后处理手段进行提名丢弃，又会导致部分信息丢失。为解决上述问题，提出一种实例图推理策略，生成时序掩码$\boldsymbol{A}\in\mathbb{R}^{\mathrm{T}}$显式引导信息聚合：

$$\mathcal{G}\rightarrow\boldsymbol{A}$$

首先初始化实例图$\mathcal{G}=\{V,E\}$，其中V表示图节点，由该片段序列中T个实例构成，$E=\Phi$。随后开始推理流程：1）节点删除，设置分类阈值θ^{vid}，对预测概率$\hat{p}$中行为类别的概率最大值小于分类阈值的实体提名，将其视为背景，并删除实例图中对应的节点；2）边构建，对于其余节点，将$\hat{p}_i$作为节点分数进行降序排序并设置相关性阈值θ^r，对于最大的节点对应实例，计算与其余节点实例的交并比，在图中将大于相关性阈值θ^r的节点与该节点相连，然后重复该过程实现对示例图的推理，得到最终的实例图$\mathcal{G}$；3）将$\mathcal{G}$中具有相连边的节点子集对应的实例进行边界加权平均融合，其权重为各个子图中节点分数与交并比乘积的归一化，得到融合实例提名集$\hat{I}'=\{(\hat{p}_i,\hat{s}_i,\hat{e}_i)\}_{i=1}^{M_G}$，其中$M_G$表示融合后的实例数量。

时序掩码生成：不同于离线视频行为检测的后处理操作，融合实例提名集在抑制低质量提名信息的同时能有效融合质量相近的提名信息。然而，多个提名间的时间区域仍然会存在重叠，为了生成有效的二值时序掩码，设计如图 2 所示的时序掩码生成方法。首先将存在区域重叠的时序提名集转换为统计分布，对于在$\hat{I}'$中的提名，将其覆盖的片段在时间维度上进行频数统计，针对所有提名完成该统计后将其转换为频率，并以此作为行为实例信息在时间上分布χ的估计。对于分布χ中的任意一个片段的概率表明了该片段属于行为实例的可靠性。设置动态阈值$\theta^{\chi}=\beta\max_t\chi(t)$，其中$\beta$表示动态阈值参数。对于大于该动态阈值的片段，将其视作有效片段。然后构建长度为T的前景序列$\boldsymbol{A}_c=\{m_1,m_2,\cdots,m_t\}$，对于序列元素$m_i$，若第$i$个片段有效，则令$m_i=1$，否则置零。

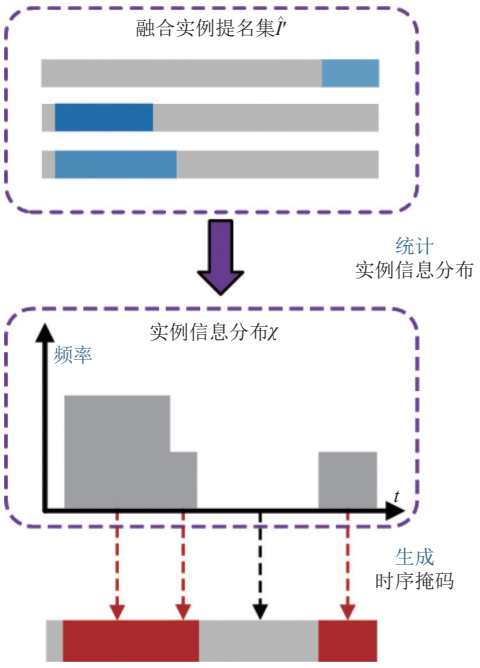

图 2 时序掩码生成示意图

以上生成的前景序列能够描述每一个片段是否属于行为实例，然而对于当前时刻片段的前背景属性未知，为了生成最终的时序掩码，设计一种前背景分数，计算方式为：

$$\mu = \max(\{\hat{p}_{0,1}, \hat{p}_{0,2}, \cdots, \hat{p}_{0,C}\})$$

依据前背景分数得到时序掩码：

$$\boldsymbol{A} = \begin{cases} A_C & \mu \geqslant 0.5 \\ 1 - A_C & \mu < 0.5 \end{cases}$$

2.4　特征融合模块

在完成实例导引生成时序掩码后，设计特征融合模块，基于实例信息引导聚合时序特征得到实例特征，构建实例分类器实现在线视频流行为检测。首先基于时序掩码将时序特征$\boldsymbol{F}^{\text{frame}}$在时序上做特征融合，得到实例特征：

$$\boldsymbol{f}^{\text{ins}} = \alpha \left(\frac{1}{\sum_{i=-T+1}^{0} A_i} \sum_{i=-T+1}^{0} A_i \boldsymbol{F}_i^{\text{frame}} + \boldsymbol{F}_0^{\text{frame}} \right)$$

式中，α为融合超参数。在此融合范式中，第一项表示按照时序掩码为导引，在时间维度上自适应聚合时序特征后，再基于时序掩码的时序取值做归一化；第二项则类似残差连接，在加速收敛的同时强化当前片段信息对实例特征的贡献，防止因时序掩码噪声导致的非相关信息对实例特征的劣化。融合超参数α保证了实例特征与时序特征在特征空间的分布同一性。最后将$\boldsymbol{f}^{\text{ins}}$输入与帧级分类器$\text{Cls}^P(\cdot)$结构相同的实例分类器$\text{Cls}^I(\cdot)$，得到最终的在线视频流行为检测结果$\hat{\boldsymbol{p}}_{\text{ins}}$，并使用交叉熵损失作为分类损失$\mathcal{L}_{\text{ins}}$。

2.5　训练与测试

在训练阶段，使用离线未裁剪视频模拟在线视频流进行训练。对于每个视频进行片段划分并完成特征提取后，设置长度为 T 的窗口对整个视频特征序列进行滑窗采样，采样出的样本作为训练样本进行计算。依据以上方法设计理论，采用以下损失函数作为目标函数进行模型训练优化：

$$\mathcal{L} = \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_r + \lambda_3 \mathcal{L}_{\text{ins}}$$

式中，$\lambda_1, \lambda_2, \lambda_3$表示多任务平衡因子。在测试时，将$\hat{p}_{\text{ins}}$作为最终行为预测概率，并参与性能指标计算。

3　实验结果及分析

3.1　实验数据集

为验证方法性能，搜集网络公开的未裁剪视频数据[19]构建数据集并进行训练与测试。所构建数据集共 20 个行为类别，每个视频平均 15 个行为实例。其次对搜集数据进行划分，其中训练集数据为 213 个视频，测试集数据为 200 个视频。

3.2　实验指标

为了验证方法的性能，使用逐帧平均精度均值 (per-frame mean average precision, mAP) 作为性能指标并与其他方法进行对比。实验收集每个片段的分类分数，并根据排序结果计算精度与召回率，然后利用精度和召回率计算每个行为类别的平均精度，并对所有类别的平均精度进行均值计算，获得最终的 mAP 结果。

3.3　实验设置

利用深度学习工作站进行实验，处理器为 Intel Core i7-13700K，GPU 型号为 NVIDIA GeForce GTX 1 080，系统环境为 Ubuntu 18.04.6 LTS；实验框架使用 Pytorch-1.10。

在视频数据预处理阶段，对每个视频以 24 FPS 的帧率进行抽帧，并以长度$l = 6$将图像帧划分为片段，即一个片段包含 6 帧视频帧，然后对每个片段利用 Denseflow 工具包计算光流图，得到光流序列。在特征提取阶段，使用在 Kinetics[20] 数据集上进行预训练的 TSN[13] 双流网络，其中光流分支使用 ResNet-200[21]，视觉分支使用 BN-Inception[22]。在网络结构部分，输入序列长度$T = 64$，映射层维度设置为 1 024，时序特征编码器中 MHA 头数为 8，前馈网络映射维度为 1 024，编码器级联数量$L = 4$。在实例导引生成器中，分类阈值$\theta_{\text{vid}} = 0.1$，相关性阈值$\theta_r = 0.4$，动态阈值参数$\beta = 0.4$。在特征融合过程中，融合参数设置为 0.5。在训练过程中，使用 AdamW 优化器进行参数优化，学习率为 0.000 1，权重衰减设置为 0.000 05，动量参数设置为 0.9，目标函数多任务平衡因子$\lambda_1, \lambda_2, \lambda_3$分别设置为 0.5, 1, 1。在后续实验中，将仅使用时序编码器进行时序编码，并在其后构建逐帧分类器的网络结构作为实验基线。

3.4　对比实验

使用以上实验设置在目标数据集上进行实验，并将实验结果分别与 CNN、RNN、Transformer 架构的视频流行为检测方法进行对比，对比结果如表 1 所示。为了公平比较，首先将提出方法与实验设置一致的先进方法进行对比，实验结果表明，本文方法能够有效对视频流中的行为进行检测。特别是与使用相同 Transformer 架构作为时序编码器的 OadTR 方法具有 65.2% 的 mAP 相比，本文方法检

测性能优于已有方法。其次，为了进行更为广泛的比较，将时序编码器进行替换，将提出的实例导引与 MiniROAD 算法中的 RNN 时序特征编码器耦合为 MR-IIG-VSAD，实验结果表明实例信息导引能够有效增强已有单一帧级架构的检测方法，且具备泛化性能。

表 1　方法间检测性能对比消融实验

方法	网络结构			mAP/%
	CNN	RNN	Transformer	
TRN[12]	√	√		62.1
IEU[23]		√		60.4
OadTR[14]	√		√	65.2
Colar[24]	√			66.9
TLS-RWKV[25]			√	71.8
MiniROAD[13]		√		71.8
MKD-CA[15]	√		√	69.9
IIG-VSAD	√		√	67.2
MR-IIG-VSAD	√	√	√	**72.4**

此外在本地环境中进行了算法效率及复杂度对比，如表 2 所示。具体为 RNN 架构下的 TRN 和 Transformer 架构下的 OadTR，其中 TRN-Step 表示保留历史信息的单步预测。结果表明本文方法在相同数据条件下，在检测效率上具有优势，有效的实例导引能够在保证检测性能的同时降低模型的宽度与深度，进而减少模型参数量与 GFLOPs，降低模型复杂度。与同为 Transformer 架构的 OadTR 相比，本文方法虽然引入了额外的实例导引分支，却实现了检测性能与检测效率的同时超越。这表明虽然 Transformer 架构的视频流行为检测方法具备全局信息提取能力，但其面向帧级的目标函数及优化策略在一定程度上抑制了模型能力。

表 2　检测效率及复杂度对比

方法	FPS	GFLOPs	参数量×10^6/个
TRN[12]	1.8	88.79	402.9
TRN-Step[12]	40.3	**1.46**	402.9
OadTR[14]	53.8	2.52	75.7
IIG-VSAD	**204.1**	2.96	**47.28**

3.5　消融实验

为了验证本文方法中各个组成部分的有效性，在目标数据集中进行了广泛的消融实验。

3.5.1　各组件的有效性

为了验证本文方法各个组件的有效性，针对组件进行消融实验。以基线、实例导引（将实例导引生成器中当前片段的分类结果$\hat{p}_0$作为检测结果）与本文方法作为实验设置，实验结果如表 3 所示。结果表明，基线仅使用与其他方法相同的逐帧分类范式仅具有 64.9% 的 mAP，说明模型对单帧信息的敏感性影响了最终的检测性能。另一方面，仅使用实例导引分支的检测性能达到 65.7%，可以看出，实例导引分支在完成实例级信息的训练优化后，能够降低模型对帧级信息的依赖，提高模型检测的实例一致性。最终利用实例信息完成特征融合后，本文方法有效地实现了两种不同层级的信息融合，得到了最优的检测性能。

表 3　组件消融实验

实验设置	mAP/%
基线	64.9
实例导引	65.7
IIG-VSAD	**67.2**

3.5.2　实例图推理的有效性

本文提出了实例图推理策略，以最大化利用预测的实例信息来生成优质的实例导引，即时序掩码。为验证该策略的有效性，设置 3 种不同的掩码生成策略，分别为：不使用实例信息（直接利用行为预测概率序列作为时序掩码）、非极大值抑制（利用非极大值抑制后剩余的实例提名生成时序掩码）以及提出的图推理策略，其消融实验结果如表 4 所示。首先不使用实例信息生成时序掩码后其检测性能达到 66.5%，这表明实例导引生成器中完成分类器和回归器的同时训练后，仅利用帧级信息进行预测也能间接提升检测性能。然而，使用与离线行为检测类似的非极大值抑制策略来滤除低质量提名，以生成时序掩码，其性能仅达到 65.4%。可以看出，虽然引入了高质量实例信息用于特征融合，但非极大值抑制这种用于离线行为检测的后处理操作丢失了大量的行为定位信息，导致了模型检测性能的劣化。最终在引入提出的图推理策略后，能够实现对实例提名信息的高质量利用，进一步提升行为模型检测能力。

表 4　实例图推理有效性实验

实验设置	mAP/%
不使用实例信息	66.5
非极大值抑制	65.4
图推理策略	**67.2**

3.5.3　特征聚合方法

为了得到辨别性的实例特征，首先利用时序掩

码自适应聚合特征，其次利用时序掩码进行特征归一化，利用类残差方式强化当前片段特征信息与实例特征的比重，最终融合超参数保证融合前后特征分布的统一。为验证以上聚合思路，进行了如表 5 所示的消融实验。其中，直接聚合表示直接利用时序掩码聚合特征，其检测性能为 66.4% mAP。事实上，在线视频流行为检测关注点仍然是当前片段的特征信息，虽然直接聚合能够从实例层面有效聚合信息，却弱化了当前片段信息对最终预测的影响。因此，在加入类残差连接后，实现了 66.7% mAP。在此基础上，通过引入融合超参数，最终使模型在目标数据集上的检测性能达到 67.2% mAP。

表 5　特征聚合方式的影响

实验设置	mAP/%
直接聚合	66.4
+类残差连接	66.7
IIG-VSAD	**67.2**

3.6　定性分析

为了进一步证明本方法的有效性，本文进行了如图 3 所示的可视化分析。具体地，将模型对视频流的推理流程中的行为预测结果进行可视化，与表 3 的实验设置类似，从上到下分别为真实标签、基线、实例导引以及提出的 IIG-VSAD。可视化结果显示，虽然帧级分类从整体性能上能够实现如表 3 实验结果所示的有效检测，但对于同一个行为实例，帧级分类的检测方式会出现如图 3 上部分所示的第 1 个实例出现的实例断连现象，这种现象中模型将同一个行为实例分割为多个实例，这是由于仅使用帧级信息监督缺失对行为整体的一致性认知导致的。另一方面，仅使用实例导引进行行为检测会出现图 3 所示的实例粘连现象（多个实例被连接成一个实例），这是由于模型过度重视实例的整体性导致的检测错误，以上现象导致设计的检测器仍然无法有效用于现实应用。而在 IIG-VSAD 中，实例信息仅用作生成用于特征融合的时序导引，同时仍然重视当前片段信息对最终检测结果的贡献，实现了更好的检测结果。由图 3 中两个视频段的可视化结果显示，本文方法能够在避免实例断联现象的同时有效区分相邻实例间的背景片段。此外，从可视化结果中可以看出，针对单个行为实例，本文方法检测实例更加完整，能够有效地辨别行为的边界。

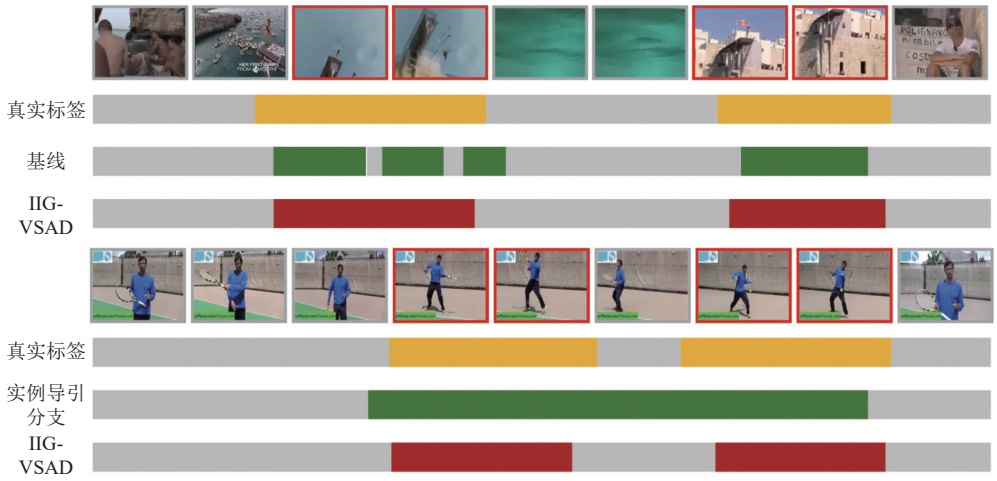

图 3　IIG-VSAD 检测结果可视化分析

4　结束语

本文提出了一种基于实例信息引导的在线视频流行为检测方法 IIG-VSAD。首先在完成视频流数据预处理以及特征提取后，构建时序特征编码器，在自注意力机制作用下提取时序片段特征。然后设计时序导引生成器，预测实例提名集，并设计实例图推理策略生成时序掩码。最终构建特征融合模块聚合实例特征并依此预测行为类别，实现在线视频流行为检测。实验证明本文方法在目标数据集上具备优秀的检测性能，并能保持高效实时运算。同时，设计了大量的消融实验，验证了本文方法细节的合理性和有效性。同时，在时序检测任务中引入实例信息进行性能增强，这个思路为后续相关研究提供了参考。

参 考 文 献

[1] SHU T M, XIE D, ROTHROCK B, et al. Joint inference of groups, events and human roles in aerial videos[C]// Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition. [S.l.]. IEEE, 2015: 4576-4584.
[2] 翟强, 程洪, 黄瑞, 等. 智能汽车中人工智能算法应用及其安全综述[J]. 电子科技大学学报, 2020, 49(4): 490-498
ZHAI Q, CHENG H, HUANG R, et al. A survey: Artificial intelligence and its security in intelligent vehicle[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(4): 490-498.
[3] MU Y, ZHANG Q, HU M, et al. Embodiedgpt: Vision-language pre-training via embodied chain of thought[EB/OL]. [2024-09-23]. https://openreview.net/pdf?id=IL5zJqfxAa.
[4] 王柏村, 薛塬, 延建林, 等. 以人为本的智能制造: 理念、技术与应用[J]. 中国工程科学, 2020, 22(4): 139-146
WANG B C, XUE Y, YAN J L, et al. Human-centered intelligent manufacturing: Overview and perspectives[J]. Strategic Study of CAE, 2020, 22(4): 139-146.
[5] SHI D F, ZHONG Y J, CAO Q, et al. ReAct: Temporal action detection with relational queries[M]//Computer Vision – ECCV 2022. Cham: Springer, 2022: 105-121.
[6] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[M]//Computer Vision – ECCV 2020. Cham: Springer, 2020: 213-229.
[7] ZHU Z, WANG L, TANG W, et al. ContextLoc++: A unified context model for temporal action localization[J]. IEEE Trans Pattern Anal Mach Intell, 2023, 45(8): 9504-9519.
[8] LIN T W, LIU X, LI X, et al. BMN: Boundary-matching network for temporal action proposal generation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 3889-3898.
[9] FOO L G, LI T J, RAHMANI H, et al. Action detection via an image diffusion process[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 18351-18361.
[10] LIU S M, ZHANG C L, ZHAO C, et al. End-to-end temporal action detection with 1B parameters across 1000 frames[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 18591-18601.
[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
[12] XU M Z, GAO M F, CHEN Y T, et al. Temporal recurrent networks for online action detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 5532-5541.
[13] AN J, KANG H, HAN S H, et al. MiniROAD: Minimal RNN framework for online action detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 10341-10350.
[14] WANG X, ZHANG S W, QING Z W, et al. OadTR: Online action detection with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2021: 7565-7575.
[15] LIU T S, LAM K M, BAO B K. A memory-assisted knowledge transferring framework with curriculum anticipation for weakly supervised online activity detection[J]. International Journal of Computer Vision, 2025, 133(4): 1940-1963.
[16] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[M]//Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 20-36.
[17] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.
[18] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS: Improving object detection with one line of code[C]// Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 5561-5569.
[19] IDREES H, ZAMIR A R, JIANG Y G, et al. The THUMOS challenge on action recognition for videos “in the wild”[J]. Computer Vision and Image Understanding, 2017, 155: 1-23.
[20] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 6299-6308.
[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.
[22] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. [S.l.]: [s.n.], 2015: 448-456.
[23] MIN S, MOON J. Information elevation network for online action detection and anticipation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2022: 2550-2558.
[24] YANG L, HAN J W, ZHANG D W. Colar: Effective and efficient online action detection by consulting exemplars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 3160-3169.
[25] ZHU Z Q, SHAO W C, JIAO D D. TLS-RWKV: Real-time online action detection with temporal label smoothing[J]. Neural Processing Letters, 2024, 56(2): 57.

责任编辑 税 红