基于时空自适应融合的双模行为识别

卿宇寒[1],高陈强[2*],谭卓林[1],刘芳岑[1]
(1. 重庆邮电大学通信与信息工程学院,重庆 400065;2. 中山大学·深圳智能工程学院,广东深圳 518107)

摘 要: 双模行为识别旨在通过学习不同数据模态间的互补信息,弥补单一模态的局限性,提升复杂场景下的行为识别性能. 现有方法通常采用独立主干网络分别提取各模态特征后再融合,但未能充分考虑模态间的语义差异(如特征不对齐),且难以有效处理模态遮挡问题,导致融合过程中易引入干扰并影响识别性能. 为此,本文提出一种基于时空自适应融合的双模行为识别方法. 具体而言,本文设计了时序关键帧选择模块,通过竞争机制突出时序关键帧;同时提出空间显著区域选择模块,自适应筛选模态间有效特征区域以抑制无关信息干扰,进而引导网络高效学习动作相关的时空特征. 此外,本文引入自蒸馏机制,结合预测分布损失和区域蒸馏损失,引导网络聚焦关键动作区域. 为进一步优化双模态特征融合效果,本文设计自适应掩码融合模块,在多头自注意力和多层感知器计算中,通过掩码过滤无效区域,降低其对特征融合的负面影响. 相比于基线方法,本文方法在 InfRA 和 NTU RGB+D 数据集上 Top-1 准确率分别提升 3.75% 和 3.49%,验证了网络能有效实现双模态特征的自适应选择与融合,提升行为识别性能.

关键词: 双模行为识别;关键帧;显著区域;自蒸馏;自适应融合
基金项目: 国家自然科学基金(No.62176035);深圳市基础研究项目(No.JCYJ20240813151216022)
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112(2025)07-2389-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250026

Bimodal Action Recognition Based on Spatiotemporal Adaptive Fusion

QING Yu-han[1], GAO Chen-qiang[2*], TAN Zhuo-lin[1], LIU Fang-cen[1]
(1. *School of Communications and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*; 2. *School of Intelligent Systems Engineering*, *Shenzhen Campus of Sun Yat-sen University*, *Shenzhen*, *Guangdong* 518107, *China*)

Abstract: Bimodal action recognition aims to enhance recognition performance in complex scenarios by leveraging complementary information across different data modalities to overcome the limitations of single-modal approaches. Existing methods typically adopt independent backbone networks to extract features from each modality separately before performing feature fusion. However, they often fail to adequately address semantic discrepancies between modalities, such as cross-modal feature misalignment and representational inconsistency, which can introduce noise during the fusion process and degrade recognition accuracy. To address these issues, this paper proposes a spatiotemporal adaptive fusion framework for bimodal action recognition. Specifically, a temporal keyframe selection module is introduced to identify and emphasize informative frames through a competitive mechanism. Simultaneously, a spatial salient region selection module adaptively filters discriminative regions across modalities, suppressing irrelevant information and guiding the network to learn more robust spatiotemporal representations. In addition, a self-distillation mechanism is employed to reinforce the network's focus on action-relevant features, incorporating both prediction distribution loss and region-level distillation loss to facilitate fine-grained feature optimization. To further improve the fusion quality, an adaptive mask fusion module is proposed, which attenuates the influence of uninformative regions by applying learnable masks within the multi-head self-attention and multi-layer perceptron computations. Experimental results on the InfRA and NTU RGB+D datasets demonstrate that the proposed method achieves Top-1 accuracy improvements of 3.75% and 3.49%, respectively, compared to baseline models, validating the effectiveness of the proposed framework in adaptively selecting and integrating bimodal features for improved action recognition.

Key words: bimodal action recognition; key frame; salient region; self-distillation; adaptive fusion
Foundation Item(s): National Natural Science Foundation of China (No.62176035); Shenzhen Fundamental Research Program (No.JCYJ20240813151216022)

收稿日期:2025-01-07;录用日期:2025-06-09;责任编辑:覃怀银
*通讯作者:高陈强

1 引言

行为识别旨在分析视频片段中的动作，近年来因其广泛的应用而备受关注，包括视频监控[1]、视频检索和人机交互等领域. 目前的主流方法主要聚焦于可见光模态，可见光模态作为行为识别的重要信息来源，具备捕捉丰富细节、颜色和纹理特征的优势，在正常光照条件下表现出色.

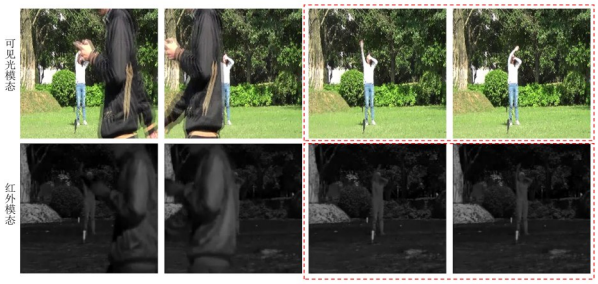
然而，随着应用场景日益复杂，单一可见光模态难以满足多样化的需求，尤其是在低光照或过曝条件下表现不佳. 相较之下，红外模态通过物体热辐射成像[2]，在低光照、雾霾或其他恶劣环境下仍然能够有效识别目标. 因此，将可见光模态的细节捕捉能力与红外模态的环境适应性相结合，可以弥补单一模态的不足，实现全天候行为识别，有效提升复杂场景下行为识别的性能.

早期研究者通过手工设计方法来提取特征，如光流直方图、密集轨迹和运动边界直方图等，并建立特征与类别的对应关系[3]，但这些方法泛化能力有限. 与之相比，基于深度学习的方法自动从原始数据中提取判别性时空特征，主流方法包括CNN类[4~7]网络、Transformer类[8~11]网络以及混合结构网络[12]. Timesformer[8]作为其中的典型代表，通过高效的时空建模，在单模态动作识别任务中表现出色. 然而Timesformer在多模态行为识别中存在模态间语义差异建模能力不足和互补特征挖掘不充分的问题，限制了其在多模态场景下的性能. 为进一步利用模态间的互补信息，近年来的研究[13~21]尝试结合多模态数据，如可见光、红外和深度等信息. 这些方法通常在后期通过特征融合或得分融合实现信息互补，或根据模态质量动态调整融合权重，以提升整体性能.

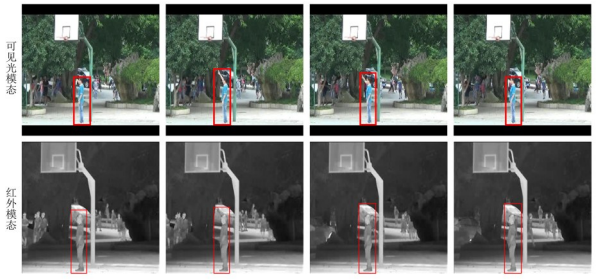
然而，由于可见光与红外模态在成像原理和视角等方面存在差异，直接使用现有融合策略会引入干扰. 如图1所示，在场景(a)中，动作主体在两种模态的部分帧中均被遮挡，同时前景中无关人物动作增加了准确识别难度. 在场景(b)中，红外模态动作主体的热辐射特征与环境背景相近，能够提供的信息有限，因此主要依赖于可见光模态的细节特征. 在场景(c)中，红外模态能够清晰捕捉动作主体，而可见光模态因动作主体与背景纹理的相似性，仅作为辅助特征参与网络学习.

基于此，在双模态行为识别任务中，某一模态在特定场景下可能因信息缺失或存在干扰而表现不佳，直接融合会影响识别性能. 为此，高效挖掘并融合双模态互补优势尤为关键. 具体而言，可见光模态擅长捕捉细节信息，而红外模态在复杂环境中表现出强适应性. 有效结合两者优势，不仅能够减少干扰，还能弥补单模态局限，提升复杂场景的行为识别性能.

同时，视频在时间维度具有天然的语义冗余性，帧间变化缓慢，仅少数帧对识别任务具有关键作用. 并且在空间维度上，单帧图像仅少数空间区域对网络学习具有实质性贡献. 因此，若能够准确定位动作主体的时序关键帧(红色虚框部分)，并剔除其余干扰帧，则可显著减少前景遮挡干扰，提高时序信息的利用效率，如图1(a)所示. 同理，在图1(b)和(c)中，当某一模态提供的关键信息有限时，网络动态筛选优先利用另一模态中显著的空间特征(红色实框部分)，并从次模态中提取补充信息，增强细节表达.



(a) 帧遮挡场景(挥手)



(b) 可见光模态主导(挥手)

(c) 红外模态主导(推)

图1 双模态行为识别复杂场景分析

本文提出基于时空自适应融合的双模行为识别方法，旨在充分利用可见光和红外模态信息，同时消除模态间干扰. 主要创新在于：(1)为充分挖掘模态间的有效特征，本文设计了时序关键帧选择模块选择具有高判别性的关键帧，同时掩码其余帧以剔除无关信息；空间显著区域选择模块在此基础上，进一步提取动作主

体区域,并对其余区域进行掩码处理,从而减少背景干扰.(2)为有效抑制掩码区域影响,本文提出自适应掩码融合模块,仅对保留特征进行交互与融合.(3)本文引入自蒸馏机制,通过教师模型提供的"软目标"(类别概率和特征)监督学生模型[22],确保网络准确定位动作主体区域.

本文在InfAR数据集[2](Infrared Action Recognition,InfAR)和NTU RGB+D[23]数据集进行大量实验.实验结果表明,本文方法在双模态行为识别任务中显著提升了性能,优于单模态以及常规融合方法.

2　相关工作

2.1　多模态融合

在行为识别中,单一模态往往难以提供充分特征,在复杂环境中表现尤为明显.为克服这一问题,多模态行为识别逐渐成为研究热点,通过融合不同模态(可见光、红外和深度等),可以弥补单一模态的局限性,提升识别性能.现有方法主要聚焦于设计有效的模态融合策略,但仍面临诸多挑战.例如,DeBoissiere等人[13]通过特征拼接实现红外和骨骼模态的联合预测,初步挖掘了模态间的互补性,但对深层次跨模态关联建模仍存在不足.Xiao等人[14]采用多层次卷积模块以增强融合效果,但其结构需要针对具体模态设计,适应性较弱.Wu等人[15]和Cheng等人[16]基于聚合模态预测得分以提升识别性能,但未能有效建模模态间的深层交互关系.为此,Song等人[17]进一步利用辅助模态优化主模态特征表示以提升整体性能.然而,当辅助模态质量较低时,融合效果可能会明显下降.

针对模态质量差异,部分研究进一步探索自适应融合机制.PDF(Predictive Dynamic Fusion)[18]通过预测模态置信度与其他模态的协同关系生成融合权重,并结合校准机制提升稳定性.TMC(Trusted Multiview Classfication)[19]基于狄利克雷分布建模模态不确定性,并借助证据理论融合多模态主观意见,使高置信模态获得更高权重.DynMM(Dynamic MultiModal fusion)[20]引入门控机制,根据输入样本动态选择模态组合与网络路径,仅激活必要的子网络以提升效率与适应性.QMF(Quality-aware Multimodal Fusion)[21]则通过能量函数估计模态质量,实现动态加权融合.尽管上述方法提升了融合灵活性,但仍难以感知模态内部的细粒度质量差异,当某一模态整体质量偏低、但局部区域包含有效信息时,容易忽略有用特征.

2.2　时空关键信息选择

在行为识别任务中,时间信息对于准确捕捉动作至关重要[24].传统的均匀采样策略难以适应动作集中发生的时序特性,易导致所选帧缺乏代表性.为此,研究者提出关键帧选择策略,增强时序信息的建模能力.例如,Korbar等人[25]将视频划分为多个片段,过滤低信息片段以聚焦于关键内容.Jiang等人[26]利用K-means聚类从各段中选取代表帧,并结合可见光和光流信息决策关键帧.Zhi等人[27]提出运动导向采样器,通过帧间特征差异与累积运动分布选择显著帧.尽管上述方法提升了时序建模效果,但依赖于复杂模块,增加了计算开销;此外,采样过程与训练解耦,易导致采样偏向特定帧,甚至错失关键帧.

在时空特征建模中,空间信息同样关键.视觉Transformer(Vision Transformer,ViT)[28]将图像划分为多个图像块,并通过自注意力机制建模区域间的长程依赖.然而,由于图像本身的高维特性,且ViT需处理所有区域,但仅有部分区域对最终预测具有实际贡献.因此,近年来研究逐步探索显著区域选择机制,以聚焦于判别性的图像区域.例如,Rao等人[29]提出分层区域筛选方法,随着网络层数加深逐步移除低信息量区域,聚焦于重要区域;Wang等人[30]采用基于锚点的选择策略,保留代表性的锚点区域;Chen等人[31]则结合时序信息,提出关键帧引导的空间裁剪策略,保留关键帧的全部区域,仅对其余帧进行裁剪.上述方法在空间显著性建模方面取得进展,但其忽略了多模态信息在时空维度上的协同作用,难以充分发挥各模态的互补优势.

2.3　自蒸馏机制

知识蒸馏通过教师模型指导学生模型学习以提升泛化能力.而自蒸馏不依赖外部教师,仅在模型内部进行知识传递.早期自蒸馏采用层级蒸馏策略,将深层分类器作为教师,浅层分类器作为学生.在训练过程中,浅层分类器不仅通过交叉熵损失拟合真实标签,还通过KL散度对齐深层输出,并借助L2损失逼近深层特征,从而提升模型判别能力[32].Ji等人[33]提出同层特征蒸馏方法,通过通道级池化建立同一深度的特征关联,并利用KL散度对齐真实标签以缓解跨层特征分布不匹配问题.进一步地,Yu等人[34]提出关系蒸馏机制,通过训练关系网络学习深层特征的类间相似性和类内区分性,引导浅层关系学习深层关系,从全局尺度优化判别能力.但上述方法设计范式受限于模态内知识迁移,未能充分挖掘跨模态的互补性与协同效应.

3　本文方法

3.1　网络结构

图2为基于时空自适应融合的双模行为识别网络结构.本文以Timesformer作为基线网络,结合时序关键帧选择、空间显著区域选择、自适应掩码融合模块和自蒸馏机制进行构建.红外支路和可见光支路输入视频序列$\boldsymbol{V}\in\mathbb{R}^{T\times H\times W}$,其中,$T$表示帧数,$H$和$W$分别表示

图像的高度和宽度. 将视频序列分割并映射为 $\boldsymbol{x}\in\mathbb{R}^{M\times N\times C}$ 的图像块,其中,M 和 N 分别表示时间和空间维度上图像块序列的长度,C 表示图像块的特征维度.

在图像块序列 $\boldsymbol{x}$ 前添加可学习的初始分类令牌,并与图像块序列共同输入网络进行特征提取. 在特征提取过程中,分类令牌逐步聚合全局语义信息. 红外与可见光支路分别生成分类令牌和图像块特征. 随后,融合红外和可见光分类令牌得到双模全局令牌,并沿时序维度拼接图像块特征得到联合特征表示 $\boldsymbol{x}'\in\mathbb{R}^{(2M)\times N\times C}$.

特征序列 $\boldsymbol{x}'$ 输入时序关键帧选择模块,该模块通过轻量级评分网络为每帧生成评分,筛选高得分帧作为关键帧,并掩码其余帧. 随后,将双模全局令牌嵌入经过时序关键帧选择模块筛选后的特征,并通过空间显著区域选择模块提取关键帧中的显著区域特征,同时掩码无关区域. 最终,特征输入自适应掩码融合模块,实现双模态特征的交互与融合,并结合自蒸馏机制进一步优化特征表示.

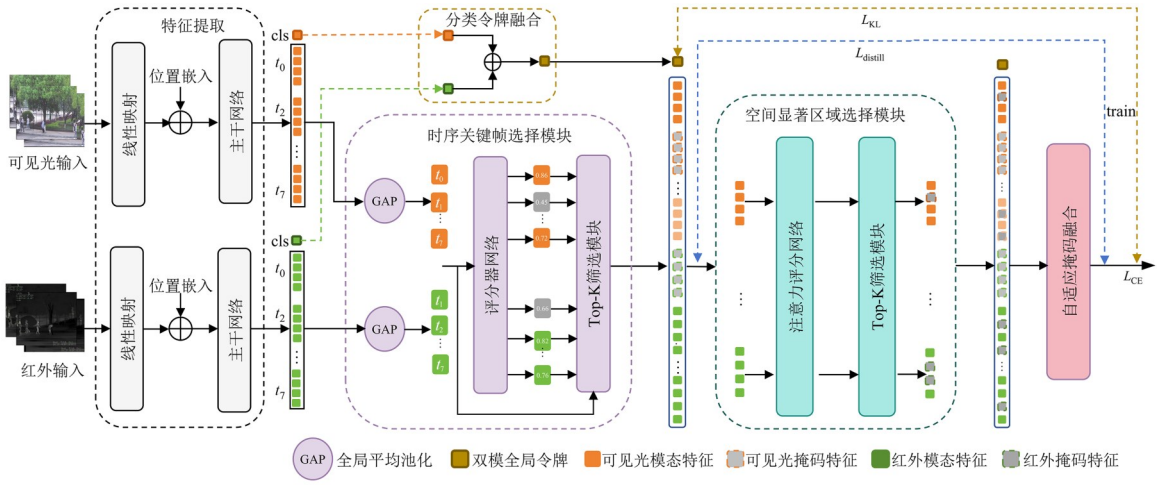


图2 基于时空自适应融合的双模行为识别网络框图

3.2 时序关键帧选择模块

现有行为识别方法通常是均匀采样视频帧并进行分类. 然而,如图1(a)所示,部分帧可能因遮挡或背景干扰而缺乏有效信息,这不仅影响网络性能,还增加了计算负担. 因此,本文提出了针对可见光和红外模态的帧间竞争选择方法,借助轻量级评分网络对每一帧的动作相关性进行评分,从而筛选出双模态中与动作类别高度相关的帧.

给定特征序列 $\boldsymbol{x}'\in\mathbb{R}^{(2M)\times N\times C}$,首先沿空间维度执行全局平均池化,得到时序特征 $\boldsymbol{x}'\in\mathbb{R}^{(2M)\times C}$,随后将其输入到全连接层 FC_1,线性映射为局部特征 $\boldsymbol{f}^{\mathrm{loc}}$,表示为

$$\boldsymbol{f}^{\mathrm{loc}}=\mathrm{FC}_1(\boldsymbol{x}')\in\mathbb{R}^{(2M)\times C'} \tag{1}$$

其中,$C'=C/2$. 为有效捕获序列的上下文信息,将局部特征 $\boldsymbol{f}^{\mathrm{loc}}$ 沿时间维度平均,得到全局特征 $\boldsymbol{f}^{\mathrm{glo}}$. 随后将局部特征 $\boldsymbol{f}^{\mathrm{loc}}$ 与全局特征 $\boldsymbol{f}^{\mathrm{glo}}$ 在通道维度上拼接,得到包含上下文信息的特征表示 $\boldsymbol{f}'$:

$$\boldsymbol{f}'=\mathrm{concat}\left(\boldsymbol{f}^{\mathrm{loc}},\boldsymbol{f}^{\mathrm{glo}}\right) \tag{2}$$

将 $\boldsymbol{f}'$ 输入全连接层 FC_2 预测每一帧的重要性 $\boldsymbol{s}'$:

$$\boldsymbol{s}'=\mathrm{FC}_2\left(\boldsymbol{f}'\right)\in\mathbb{R}^{(2M)\times 1} \tag{3}$$

对帧得分 $\boldsymbol{s}'$ 进行归一化处理,表示为

$$\boldsymbol{s}=\frac{\boldsymbol{s}'-\min(\boldsymbol{s}')}{\max(\boldsymbol{s}')-\min(\boldsymbol{s}')} \tag{4}$$

随后,生成时序帧标记向量 $\boldsymbol{M}_t\in\mathbb{R}^{2M\times 1}$,其中 $2M$ 表示双模态的总帧数. 具体地,根据归一化得分 $\boldsymbol{s}$,首先选取得分前 k_1 比例帧作为关键帧,并在 $\boldsymbol{M}_t$ 对应的索引位置生成关键帧标记;在排除关键帧后的剩余帧中,再选取得分前 k_2 比例帧作为次关键帧,并在 $\boldsymbol{M}_t$ 对应的索引位置生成次关键帧标记;其余帧则作为背景. 最终的时序帧标记向量将结合空间显著区域选择模块,进一步筛选显著区域.

3.3 空间显著区域选择模块

为消除无关区域干扰,充分利用模态内及跨模态信息,本文基于多头自注意力机制(Multi-Head Self-Attention,MHSA)实现空间显著区域选择,自适应选择红外和可见光模态中的显著区域. 具体来说,网络生成所有层的注意力权重表示为

$$\boldsymbol{A}_l=\left[\boldsymbol{a}_l^1,\boldsymbol{a}_l^2,\boldsymbol{a}_l^3,\cdots,\boldsymbol{a}_l^{N_{\mathrm{h}}}\right],\ l\in 1,2,\cdots,L \tag{5}$$

其中,L 表示网络层数,N_{h} 表示自注意力头数量. 为精确量化空间区域的相对重要性,采用跨层注意力乘积机制:

$$\boldsymbol{A}=\prod_{l=1}^{L}\boldsymbol{A}_l \tag{6}$$

该机制捕获了信息从低层到高层的传播过程,有效缓解模型层数加深时原始注意力权重不准确问

题[35],更准确地表示不同区域的相关性.

为高效处理汇总后的注意力信息$\boldsymbol{A}$,本文通过均值聚合各注意力头分数得到$\boldsymbol{A}_{\text{score}}$,衡量不同区域的重要性,表示为

$$\boldsymbol{A}_{\text{score}}=\left(\frac{1}{N_{\text{h}}}\sum_{i=1}^{N_{\text{h}}}\boldsymbol{A}\right) \tag{7}$$

随后,生成掩码标记向量$\hat{\boldsymbol{M}}_t\in\mathbb{R}^{(2M)\times N}$,在时序帧标记$\boldsymbol{M}_t$基础上,结合空间区域得分$\boldsymbol{A}_{\text{score}}$选择显著区域:

$$\hat{\boldsymbol{M}}_t=\psi\left(\text{top}_{\text{s}}\left(\boldsymbol{M}_t,\boldsymbol{A}_{\text{score}}\right)\right),\ 1\leqslant t\leqslant 2M \tag{8}$$

其中,$\hat{\boldsymbol{M}}_t$表示第t帧的空间掩码标记,top_{s}根据得分$\boldsymbol{A}_{\text{score}}$按($r_1$或$r_2$)比例选择显著区域,标记函数$\psi$将前$r_1$或$r_2$个区域标记值设为1,其余区域标记为0. 关键帧保留显著区域的比例较大r_1,以捕获丰富的空间信息;次关键帧保留显著区域的比例相对较小r_2以减少冗余特征.

3.4　自适应掩码融合模块

本文提出自适应掩码融合模块(见图3),通过掩码局部模态内注意力、掩码全局跨模态注意力以及掩码多层感知器,有效地抑制掩码区域干扰并融合模态间互补信息. 具体而言,空间显著区域选择模块根据输入生成不同数量掩码. 如果直接丢弃掩码区域会导致同一批次中有效区域数量不一致,后续计算难以并行处理. 若简单将掩码区域元素置0,这些零元素在softmax计算中被映射为一个小于1的正值,从而对未掩码区域造成干扰,影响融合效果. 因此,本文在多头自注意力计算过程中,将掩码区域的值设为负无穷,确保掩码区域对显著区域的影响被有效抑制,即:

$$\hat{\boldsymbol{M}}=\hat{\boldsymbol{M}}_t\hat{\boldsymbol{M}}_t^{\text{T}} \tag{9}$$

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=\boldsymbol{x}'\boldsymbol{W}_q,\boldsymbol{x}'\boldsymbol{W}_k,\boldsymbol{x}'\boldsymbol{W}_v \tag{10}$$

$$\boldsymbol{P}_{ij}=\begin{cases}-\infty, & \text{if } \hat{\boldsymbol{M}}_{ij}=0\\ \dfrac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}, & \text{else}\end{cases} \tag{11}$$

$$\text{Mask\_Attn}\left(\boldsymbol{x}',\hat{\boldsymbol{M}}\right)=\left(\text{softmax}\left(\boldsymbol{P}\right)\odot\hat{\boldsymbol{M}}\right)\boldsymbol{V} \tag{12}$$

其中,$\boldsymbol{Q}$为查询向量,$\boldsymbol{K}$为键向量,$\boldsymbol{V}$为值向量. 当$\hat{\boldsymbol{M}}_{ij}$对应图像块掩码标记非零时,保留该处注意力值以充分利用该区域特征.

为进一步增强局部特征的语义表达能力,本文引入掩码模态内自注意力计算,对模态内特征进行细粒度建模. 经过上述模块处理后,特征序列为

$$\boldsymbol{x}_{\text{f}}=\left\{\boldsymbol{x}_{\text{cls}},\boldsymbol{x}_1^{\text{s}},\boldsymbol{x}_2^{\text{s}},\cdots,\boldsymbol{x}_T^{\text{s}}\right\},\boldsymbol{x}_{\text{cls}}\in\mathbb{R}^{1\times 1\times C},\boldsymbol{x}_i^{\text{s}}\in\mathbb{R}^{1\times(T\times P)\times C} \tag{13}$$

其中,$\boldsymbol{x}_{\text{cls}}$表示分类令牌,$\boldsymbol{x}_i^{\text{s}}$表示帧空间特征,$T$为帧数,$P$为图像块数. 首先,在时间维度上复制分类令牌$\boldsymbol{x}_{\text{cls}}$,并将其拼接至每帧特征前部,使每帧拥有独立的帧内分类令牌,用于模态内语义信息建模,表示为

$$\boldsymbol{x}_{\text{local}}=\text{concat}\left(\boldsymbol{x}_{\text{cls}},\boldsymbol{x}_i^{\text{s}}\right)\in\mathbb{R}^{T\times(1+P)\times C} \tag{14}$$

在掩码标记$\hat{\boldsymbol{M}}$的引导下,帧内分类令牌通过式(12)建模帧内显著区域信息,增强局部特征表达能力. 随后,跨关键帧地融合各帧内分类令牌,形成统一的局部分类令牌表示$\boldsymbol{x}_{\text{local}}^{\text{cls}}$,有效学习各帧内不同区域的细粒度差异,增强了对局部细节的判别能力. 最后,将$\boldsymbol{x}_{\text{local}}^{\text{cls}}$与帧空间特征$\boldsymbol{x}_i^{\text{s}}$拼接,生成掩码模态内融合特征$\boldsymbol{x}_{\text{f}}'$.

基于更新后的特征序列,利用掩码标记$\hat{\boldsymbol{M}}$引导多头注意力计算生成跨模态的全局特征表示:

$$\boldsymbol{x}_{\text{f}}'=\text{Mask\_Attn}\left(\boldsymbol{x}_{\text{f}}',\hat{\boldsymbol{M}}\right) \tag{15}$$

该方法有效建模模态间的长程依赖关系,促进可见光与红外模态的特征交互与融合,建立跨模态全局语义关联. 随后,特征输入掩码多层感知器(Multi-Layer Perceptron,MLP),在通道维度实现特征转换与增强. 为抑制掩码区域对保留区域的干扰,在多层感知器计算前先执行掩码处理.

$$\boldsymbol{x}_{\text{f}}'=\boldsymbol{x}_{\text{f}}'\odot\hat{\boldsymbol{M}} \tag{16}$$

3.5　自蒸馏联合损失

本文通过交叉熵损失函数优化分类性能:

$$L_{\text{CE}}=-\sum_{i=1}^{K}y\left(\boldsymbol{x}_i\right)\log\left(\tilde{y}\left(\boldsymbol{x}_i\right)\right) \tag{17}$$

其中,K表示类别数,$y\left(\boldsymbol{x}_i\right)$和$\tilde{y}\left(\boldsymbol{x}_i\right)$分别表示样本的真实概率和预测概率.

为提高空间显著区域选择模块的精确性,本文将基线网络输出的特征作为教师监督,将自适应掩码融合模块输出的特征作为学生特征,通过自蒸馏机制优化学生特征对显著区域的聚焦能力. 区域蒸馏损失函数L_{distill}为

$$L_{\text{distill}}=\frac{1}{\sum_{t=1}^{2M}\sum_{i=1}^{N}\hat{\boldsymbol{M}}_t^i}\sum_{t=1}^{2M}\sum_{i=1}^{N}\hat{\boldsymbol{M}}_t\left(\boldsymbol{p}_i-\boldsymbol{p}_i'\right)^2 \tag{18}$$

其中,$\boldsymbol{p}_i$和$\boldsymbol{p}_i'$分别表示教师和学生模型保留的区域特征. 此外,通过最小化KL散度约束,减小教师和学生模型的分类预测差异:

$$L_{\text{KL}}=\sum_{i=1}^{K}y(\boldsymbol{x}_i)\log\frac{y\left(\boldsymbol{x}_i\right)}{y'\left(\boldsymbol{x}_i\right)} \tag{19}$$

其中,$y\left(\boldsymbol{x}_i\right)$和$y'\left(\boldsymbol{x}_i\right)$分别表示教师和学生模型的类别

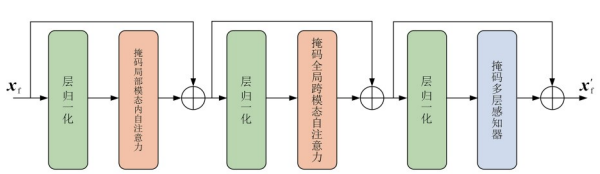


图3　自适应掩码融合模块示意图

概率.

综上所述,训练阶段的总损失函数为

$$L=L_{\mathrm{CE}}+\lambda\times L_{\mathrm{distill}}+(1-\lambda)\times L_{\mathrm{KL}} \tag{20}$$

其中,参数λ用于平衡区域蒸馏损失与预测分布损失.

4 实验分析

4.1 数据集与评估指标

为验证本文方法的有效性,在InfRA和NTU RGB+D公开数据集进行大量实验. 下面简要介绍这2个数据集.

InfRA数据集包含红外和可见光模态,涵盖12个动作类别,其中可见光分辨率为720×480,红外分辨率为293×256,提供了960组训练样本和240组测试样本. 视频时长为5~10 s,涵盖了不同季节、复杂背景、遮挡情况以及视角变化等因素,提供了丰富的行为识别场景条件.

NTU RGB+D数据集包含可见光、深度图、骨架和红外模态,包含60个类别,共56 880个样本. 由40名年龄、性别和身高各异的参与者完成,通过3台摄像头采集,丰富了视角和场景的多样性. 本文使用了其中的可见光和红外模态,可见光分辨率为1 920×1 080,红外分辨率为512×424. 并遵循官方数据集的划分方式,将摄像头1的样本用于测试,摄像头2和3的样本用于训练.

本文采用Top-1与Top-5准确率评估模型性能,两者结合以全面评估模型性能. Top-1准确率衡量模型预测的首选类别是否与真实标签一致,作为单标签分类的核心指标;Top-5准确率则检验真实标签是否位于预测的前5候选类别中,反映模型在类别相似情况下的表现.

4.2 网络细节

本文输入包括可见光和红外模态视频. 在数据采样阶段,从视频中选取8个片段,并从每个片段中随机采样一帧. 在数据预处理阶段,训练样本首先被随机缩放至260~320的范围,然后随机裁剪为224×224的固定尺寸,并以50%的概率水平翻转以增加数据的多样性. 在训练过程中,采用混合精度策略,并使用SGD优化器更新参数. 学习率初始化为0.005,动量系数设为0.9,结合Nesterov算法加速梯度下降.模型共训练100轮次,在第35轮和第70轮时衰减学习率,衰减因子为0.1.

4.3 对比实验

目前,双模行为识别研究主要集中在可见光、深度和骨骼模态,而针对可见光与红外模态的研究相对较少,现有成熟且通用的方法也较为有限. 因此,本文选择了具有代表性的通用行为识别模型作为对比方法,涵盖以下几类:基于CNN(Convolutional Neural Network)类行为识别方法TSM(Temporal Shift Module)[4]、SlowFast[5]、CSN(Channel-Separated convolutional Networks)[6]和TaNet(Temporal Adaptive Module)[7];基于Transformer类行为识别方法TimeSFormer[8]、VideoSwin[9]、MViT[10]和轻量化视觉转换器Hiera[11],以及混合结构的时空焦点调制Video FocalNets[12]. 这些方法代表了不同范式的行为识别模型,确保了对比实验的全面性.

上述方法均基于单一模态,为确保实验对比的一致性,本文将每种方法扩展为双支路结构,分别提取可见光和红外特征,并通过特征相加方式进行融合,并送入分类图进行预测. 实验结果见表1. 相较于基线方法,本文方法在InfRA数据集上的Top-1和Top-5准确率分别提升了3.75%和1.25%,在NTU RGB+D数据集上的Top-1和Top-5准确率分别提升了3.49%和1.74%,这表明本文方法能有效利用双模态信息,提高行为识别的准确性.

不同网络框架性能分析表明,CNN类方法在Top-5识别率上表现较好,但其在Top-1识别率上存在局限,说明其在决策阶段的信息利用仍有不足. 这主要是由于CNN的局部特征建模机制,在应对动作类别间局部相似性时,容易出现混淆,误分类集中于相似类别. 相比之下,Transformer类方法通过全局上下文建模降低对局部细节的依赖,在相似类别间展现出更强的判别能力,使其在Top-1识别率上表现更优,体现其精细化分类优势. 因此,Transformer结构在复杂场景中具有更强大的特征建模能力,整体表现优于CNN. 而Hybrid结构尝试结合CNN的计算效率和Transformer的建模能力,旨在平衡模型性能与计算开销,但实际增益有限.

从计算复杂度分析,尽管Transformer结构通常伴随着较高的计算开销,本文方法与基线方法对比,运算量增加了33 G Flops,参数量增加了8 M,却在Top-1精度上取得了显著提升. 这表明本文方法在适度增加计算成本时,有效利用了可见光与红外模态的互补信息,实现了更高的精度提升,进一步验证了其在复杂识别场景中的有效性.

表2对比了基线网络不同融合策略的性能影响. 单模态特征虽能提供判别信息,但由于缺乏多模态互补特征支持,识别精度受限. 简单融合(如平均、拼接)虽能在一定程度上提升准确率,但由于未能有效捕捉模态间的复杂关联,性能增益仍面临瓶颈. 相比之下,PDF[18]、TMC[19]和DynMM[20]采用模态级动态融合方式,根据一定策略为整个模态赋予权重,动态调整模态间的融合比例. 该方法虽能学习模态重要性,却无法区分模态内部的质量差异. 当某模态整体置信度较高时,其局部低质量特征仍会导致干扰;而低置信度模态中的有效局部信息则被忽视,限制了模型判别性上限.

表1　不同方法在InfRA和NTU RGB+D数据集的性能分析

实验方法		InfRA		NTU RGB+D		运算量/G	参数量/M
		Top-1/%	Top-5/%	Top-1/%	Top-5/%		
CNN	TSM[4]	49.58	92.50	73.94	96.09	66.3	23.5
	SlowFast[5]	58.33	93.33	81.31	98.08	48.8	61.9
	CSN[6]	74.58	93.75	76.59	97.22	42.7	32.2
	TaNet[7]	82.50	**98.75**	82.48	**99.17**	126	89.8
Hybrid	VideoFocalNets[12]	86.25	96.25	85.73	98.54	221	157
Transformer	VideoSwin[9]	83.33	95.83	84.75	98.75	302	196
	MViT[10]	84.58	97.92	83.59	97.95	362	213
	TimeSFormer[8]	85.42	97.08	83.18	97.33	327	101
	Hiera[11]	87.50	97.08	86.52	98.99	203	115
	本文方法	**89.17(+3.75)**	98.33(+1.25)	**86.67(+3.49)**	99.07(+1.74)	360(+33)	109(+8)

注:加粗数据最优结果,下划线数据表示基线方法.

表2　不同融合方法对比分析　　单位:%

方法	Top-1	Top-5
红外	77.50	96.67
可见光	81.67	97.08
平均	84.58	97.50
拼接	83.33	97.50
TMC[19]	84.17	98.33
DynMM[20]	85.42	97.75
PDF[18]	87.92	97.92
本文方法	**89.17**	**98.33**

注:加粗数据为最优结果.

表3　模块有效性消融实验

时空选择模块	掩码融合模块	自蒸馏	Top-1/%	Top-5/%
			85.42	97.08
√			86.67(+1.25)	97.50(+0.42)
	√		87.08(+1.66)	**98.75(+1.67)**
		√	86.25(+0.83)	96.67(-0.41)
√		√	87.50(+2.08)	97.92(+0.84)
	√	√	87.50(+2.08)	97.92(+0.84)
√		√	87.92(+2.50)	98.33(+1.25)
√	√	√	**89.17(+3.75)**	98.33(+1.25)

注:加粗数据为最优结果.

本文采用实例级动态融合策略,不仅关注模态间的互补关系,还深入模态内部,对不同时序帧和空间区域进行质量评估和选择,主动丢弃信息量较低部分. 与模态级融合方式不同,本文方法通过精细的显著特征筛选突出关键帧和显著区域,减少冗余信息并融合,有效提升了识别精度.

4.4　消融实验

本文以Timesformer为基线网络,结合时序关键帧选择、空间显著区域选择、自适应掩码融合模块与自蒸馏机制,在InfRA数据集进行了全面消融实验以验证模块有效性. 此外,本文分析了模块参数变化对模型性能的影响.

4.4.1　模块有效性消融实验

本节对各模块的有效性进行实验分析,结果见表3. 在实验中,将时序关键帧和空间显著区域选择模块作为整体分析,主要基于以下2点:一是两者均通过筛选关键信息协同提升模型性能;二是时空信息对行为识别至关重要,合并分析有助于全面评估综合效果,避免片面结论.

实验表明,时序关键帧与空间显著区域选择模块有助于模型聚焦于更具判别力的时空特征,从而提升识别性能. 自适应掩码融合模块增强了双模态间的特征交互,改善了模态间的互补性利用. 自蒸馏机制则通过引导特征学习过程,使模型在识别精度上进一步优化.

更重要的是,当所有模块协同工作时,Top-1准确率提升至89.17%,显著优于单一模块或部分模块的组合. 这充分表明各模块存在互补性与协同作用,有力支撑了本文方法在双模态行为识别任务中的有效性.

4.4.2　时序关键帧和空间显著区域选择模块消融实验

本文提出时序关键帧和空间显著区域选择模块,自适应筛选模态中的关键动作区域以减少冗余信息. 如表4所示,通过调整时序关键帧保留比例(k_1、k_2)和空间显著区域保留比例(r_1、r_2),有效提升双模态行为识别性能.

从时序关键帧角度分析,对比实验1和实验4,在相同的显著空间区域保留比例下,实验4保留的关键帧比例较低,但其Top-1准确率达到89.17%,优于实验1. 这表明,适度减少冗余帧有助于去除无关信息,提升模型性能.

从空间显著区域角度分析,在相同的关键帧比例下,实验3、实验2和实验1逐步增加显著区域的保留比

表4 时序关键帧和空间显著区域选择比例消融实验

实验	时序关键帧选择比例		空间显著区域选择比例		Top-1/%	Top-5/%
	k_1	k_2	r_1	r_2		
1	0.375	0.125	0.60	0.35	87.08	97.50
2			0.45	0.25	85.83	97.95
3			0.30	0.15	86.25	98.33
4	0.25	0.25	0.60	0.35	**89.17**	98.33
5			0.45	0.25	87.08	98.33
6			0.30	0.15	85.83	97.92
7	0.125	0.375	0.60	0.35	86.67	**98.96**
8			0.45	0.25	86.25	98.75
9			0.30	0.15	84.58	98.33

注:加粗数据为最优结果.

例,Top-1准确率呈现出先降后升的趋势. 实验3中,显著区域保留比例较低,导致关键特征未能充分覆盖,性能较低;实验2中,虽然引入了部分有效特征,但伴随更多的干扰信息,使得有效特征的增益不足以抵消干扰特征的负面影响,导致性能下降. 实验1中,显著区域比例进一步的扩大使模型能够更全面地捕捉关键信息,从而提升了分类性能. 这表明显著区域的选取需要在覆盖判别性特征和抑制冗余干扰之间取得适当平衡,以实现更优的性能提升.

此外,实验9中,由于关键帧和显著区域保留比例设置最低,模型性能明显低于基线方法. 这表明,过度压缩时序和空间信息会造成大量关键信息丢失,影响识别性能.

4.4.3 自适应掩码融合模块消融实验

为高效融合双模态特征,本文设计了基于掩码引导的自适应特征融合模块,并通过消融实验验证其有效性,见表5. 传统Transformer编码器在复杂行为识别任务中表现不佳,主要原因是在多头自注意力(MHSA)和多层感知器(MLP)计算过程中,掩码区域干扰了显著区域的特征交互,削弱了融合效果.

表5 自适应融合模块掩码策略消融实验

Masked MHSA	MaskedMLP	Top-1/%	Top-5/%
		86.25	97.92
√		87.92	98.33
	√	86.67	**98.75**
√	√	**89.17**	98.33

注:加粗数据为最优结果.

为消除这一负面影响,本文在MHSA和MLP中分别引入了掩码标记. 从表5分析可得,仅在MHSA引入掩码时,模型的Top-1精度提升至87.92%,优于仅在MLP中引入掩码时的精度86.67%. 这表明MHSA中的掩码引入能有效滤除背景干扰,聚焦关键行为区域,从而增强全局特征建模. 相比之下,在MLP中的掩码虽有助于细化局部特征,但由于对特征位置间的交互影响较小,提升效果有限. 当MHSA和MLP同时引入掩码时,模型的Top-1精度达到89.17%,性能最优. 这表明结合MHSA的全局特征筛选优势与MLP的局部特征细化能力,模型抑制无关区域干扰,同时准确学习与行为相关的关键特征,从而显著提升整体性能.

4.4.4 自蒸馏机制消融实验

为验证自蒸馏机制的作用,本文设计了7组消融实验,参数λ分别设置为0、0.1、0.3、0.5、0.7、0.9和1,评估其对模型识别性能的影响. 如图4所示,随着λ的增加,Top-1准确率先增后减,当$\lambda=0.1$时,达到最高值89.17%. 这表明,区域蒸馏损失L_{distill}与预测分布损失L_{KL}对网络的学习均具有重要作用,而适当的权重比例能有效引导网络聚焦于动作主体相关区域,从而提升分类性能.

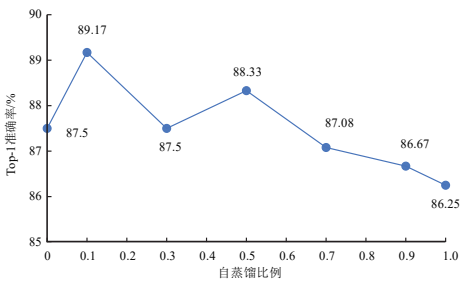


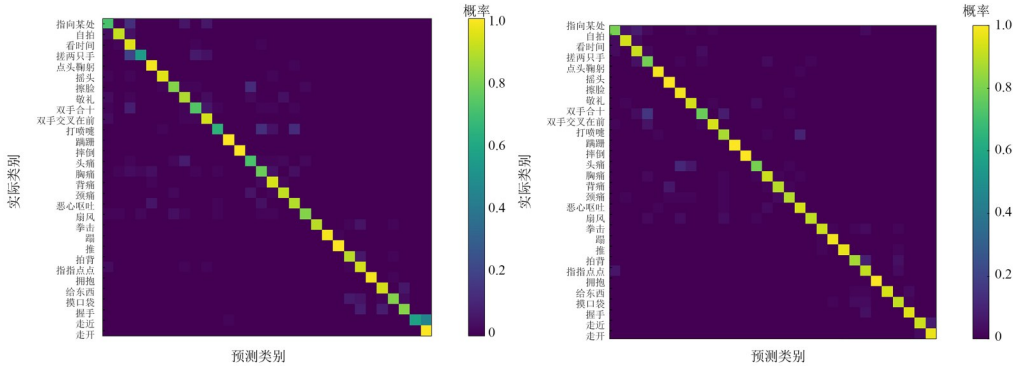
图4 自蒸馏损失权重λ消融实验

4.5 可视化分析

图5展示了基线方法Timesformer与本文方法在NTU RGB+D数据集第30~60类样本的分类结果混淆矩阵. 相比基线方法,本文方法的混淆矩阵对角线更加明亮,表明各类别的预测准确率有所提升. 具体而言,基线方法在“揉两只手”和“走开”这2类动作上误判率较高. 本文方法通过时序关键帧选择模块去除了无关帧,确保动作方向的正确捕捉;同时,空间显著区域选择模块聚焦于动作细节差异,有效区分相似动作,降低了这两类的误分类.

为直观体现本文方法的性能提升,图6和图7分别在InfRA与NTU RGB+D数据集上,通过叠加注意力掩码可视化关键帧与显著区域的选择,其中关键帧中的显著区域被保留以清晰标识动作主体区域,其余区域则被掩码以降低干扰. 每个类别包含两行图像序列,分别对应可见光与红外模态.

图6(a)和图6(b)展示了InfRA数据集中“打架”和“挥手”场景的可视化分析. 图6(a)为“打架”场景,可见光模态在3~4帧呈现动作主体逐渐靠近的过程,而红外模态中7~8帧为动作主体逐渐分开的过程,这些帧与

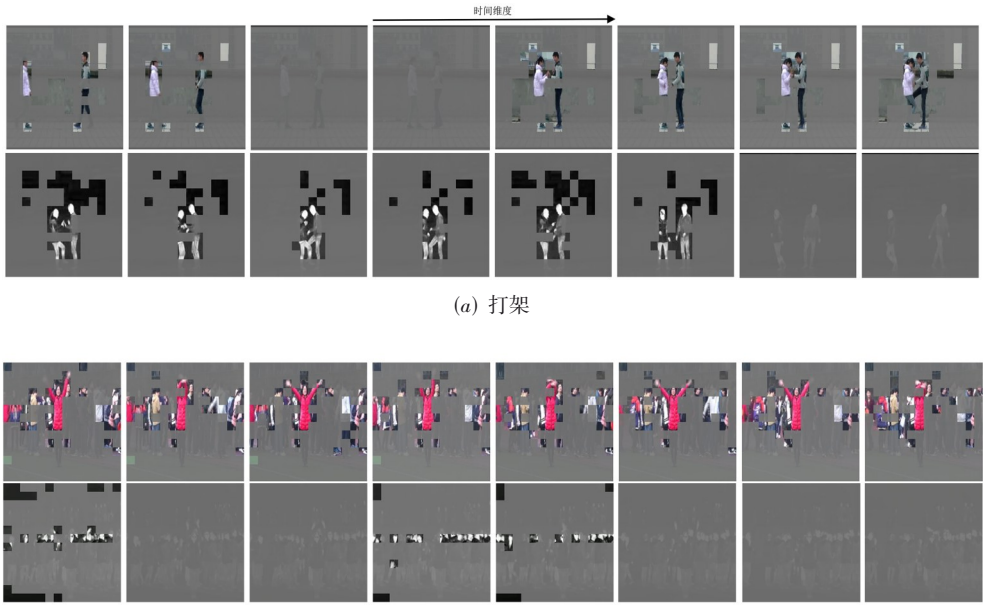


(*a*) 基线方法　　(*b*) 本文方法

图5　不同方法在NTU RGB+D数据集上第30~60类样本混淆矩阵

“打架”动作相关的信息较少,因此被掩码处理,确保网络聚焦于类别关键帧. 图6(*b*)为“挥手”场景,可见光模态中动作主体穿着特征提供了丰富的色彩和纹理信息,有助于网络精准识别动作主体,因此有效保留了主体区域;红外模态中由于背景的热辐射干扰,难以提取有效的区分特征,最终仅保留了少量区域.

图7(*a*)和图7(*b*)展示了NTU RGB+D数据集上“走开”和“自拍”场景的可视化结果. 图7(*a*)为“走开”场景,可见光与红外模态在1~3帧均呈现出动作主体靠近的趋势,这种视觉特征易被误判为“走向”动作类别. 因此,本文通过掩码机制有效屏蔽了这些干扰帧,重点关注后续帧中“走开”动作的显著特征,从而减少了误分类. 图5中的混淆矩阵进一步验证了该模块在纠正“走向”和“走开”的分类错误的有效性. 图7(*b*)为“自拍”场景,可见光模态捕捉了动作细节特征,红外模态有效抑制背景干扰,2种模态优势互补提升了识别准确率.



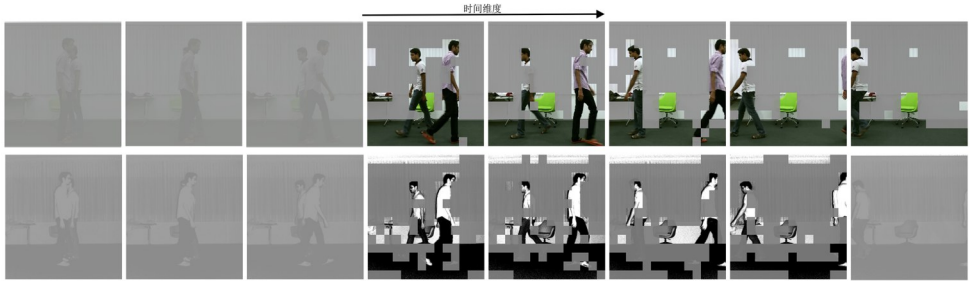
(*a*) 打架

(*b*) 挥手

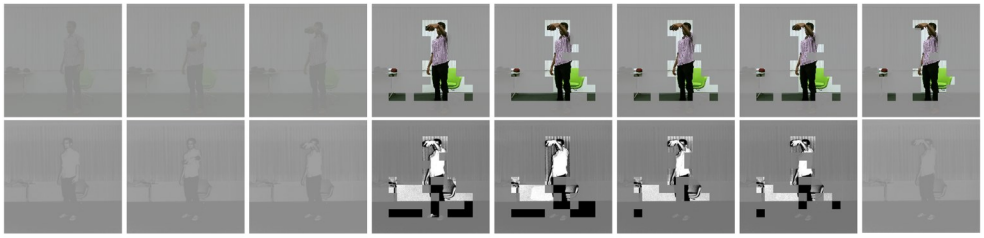
图6　InfRA样本序列可视化

5　结论

本文提出了一种基于时空自适应融合的双模态行为识别方法,旨在充分利用红外和可见光模态的有效信息,提升行为识别的准确性. 方法分为特征选择和特征融合2部分. 特征选择部分包括时序关键帧选择和空间显著区域选择模块,其中时序关键帧选择模块用于提取时序中关键信息帧,空间显著区域选择模块用于自适应筛选关键帧中变化显著的区域,引入了自蒸馏机制辅助网络定位关键区域,提升特征选择的效果. 特征融合部分设计了自适应掩码融合模块,在有效融合红外和可见光模态信息的同时,通过掩码标记抑制



(*a*) 走开



(*b*) 自拍

图7 NTU RGB+D样本序列可视化

无关信息的干扰,进一步提高融合的鲁棒性与准确性.在InfRA和NTU RGB+D数据集上的实验结果表明,本文方法在双模态行为识别任务中表现出较高的分类性能.在未来研究中,将探索逐步丢弃被掩码区域的特征,以解决当前特征融合过程中掩码区域仍参与计算的问题,实现更高效的模态交互.

参考文献

[1] ELHARROUSS O, ALMAADEED N, AL-MAADEED S. A review of video surveillance systems[J]. Journal of Visual Communication and Image Representation, 2021, 77: 103116.

[2] GAO C Q, DU Y H, LIU J, et al. InfAR dataset: Infrared action recognition at different times[J]. Neurocomputing, 2016, 212: 36-47.

[3] 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)

[4] LIN J, GAN C, HAN S. TSM: Temporal shift module for efficient video understanding[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 7082-7092.

[5] FEICHTENHOFER C, FAN H Q, MALIK J, et al. SlowFast networks for video recognition[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 6201-6210.

[6] TRAN D, WANG H, FEISZLI M, et al. Video classification with channel-separated convolutional networks[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 5551-5560.

[7] LIU Z Y, WANG L M, WU W, et al. TAM: Temporal adaptive module for video recognition[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 13688-13698.

[8] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding? [EB/OL]. (2021-06-09)[2024-12-16]. https://arXiv.org/abs/2102.05095.

[9] LIU Z, NING J, CAO Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 3202-3211.

[10] LI Y H, WU C Y, FAN H Q, et al. MViTv2: Improved multiscale vision transformers for classification and detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 4794-4804.

[11] RYALI C, HU Y T, BOLYA D, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles[EB/OL]. (2023-06-01) [2024-12-16]. https://arXiv.org/abs/2306.00989.

[12] WASIM S T, KHATTAK M U, NASEER M, et al. Vid-

eo-FocalNets: Spatio-temporal focal modulation for video action recognition[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 13732-13743.

[13] DE BOISSIERE A M, NOUMEIR R. Infrared and 3D skeleton feature fusion for RGB-D action recognition[J]. IEEE Access, 2020, 8: 168297-168308.

[14] XIAO X J, REN Z L, LI H, et al. SlowFast multimodality compensation fusion swin transformer networks for RGB-D action recognition[J]. Mathematics, 2023, 11(9): 2115.

[15] WU H B, MA X, LI Y B. Spatiotemporal multimodal learning with 3D CNNs for video action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1250-1261.

[16] CHENG J, REN Z L, ZHANG Q S, et al. Cross-modality compensation convolutional neural networks for RGB-D action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1498-1509.

[17] SONG S J, LIU J Y, LI Y H, et al. Modality compensation network: Cross-modal adaptation for action recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 3957-3969.

[18] CAO B, XIA Y, DING Y, et al. Predictive dynamic fusion[EB/OL]. (2024-06-07) [2025-03-31]. https://arxiv.org/abs/2406.04802.

[19] HAN Z B, ZHANG C Q, FU H Z, et al. Trusted multi-view classification with dynamic evidential fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2551-2566.

[20] XUE Z H, MARCULESCU R. Dynamic multimodal fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2023: 2575-2584.

[21] ZHANG Q Y, WU H T, ZHANG C Q, et al. Provable dynamic fusion for low-quality multimodal data[C]//Proceedings of the 40 th International Conference on Machine Learning. Honolulu: PMLR, 2023: 41753-41769.

[22] 郑云飞, 王晓兵, 张雄伟, 等. 基于金字塔知识的自蒸馏HRNet目标分割方法[J]. 电子学报, 2023, 51(3): 746-756.
ZHENG Y F, WANG X B, ZHANG X W, et al. The self-distillation HRNet object segmentation based on the pyramid knowledge[J]. Acta Electronica Sinica, 2023, 51(3): 746-756. (in Chinese)

[23] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1010-1019.

[24] 柯逍, 缪欣, 郭文忠. 基于时空交叉感知的实时动作检测方法[J]. 电子学报, 2024, 52(2): 574-588.
KE X, MIAO X, GUO W Z. Real-time action detection based on spatio-temporal interaction perception[J]. Acta Electronica Sinica, 2024, 52(2): 574-588. (in Chinese)

[25] KORBAR B, TRAN D, TORRESANI L. SCSampler: Sampling salient clips from video for efficient action recognition[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 6231-6241.

[26] JIANG M, PAN N, KONG J. Spatial-temporal saliency action mask attention network for action recognition[J]. Journal of Visual Communication and Image Representation, 2020, 71: 102846.

[27] ZHI Y, TONG Z, WANG L M, et al. MGSampler: An explainable sampling strategy for video action recognition[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 1493-1502.

[28] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2024-12-16]. https://arXiv.org/abs/2010.11929.

[29] RAO Y M, ZHAO W L, LIU B L, et al. DynamicViT: Efficient vision transformers with dynamic token sparsification[EB/OL]. (2021-10-26) [2024-12-16]. https://arXiv.org/abs/2106.02034.

[30] WANG J K, YANG X T, LI H D, et al. Efficient video transformers with spatial-temporal token selection[M]// Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 69-86.

[31] CHEN L, TONG Z, SONG Y B, et al. Efficient video action detection with token dropout and context refinement[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 10354-10365.

[32] ZHANG L F, SONG J B, GAO A N, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 3712-3721.

[33] JI M, SHIN S, HWANG S, et al. Refine myself by teaching myself: Feature refinement via self-knowledge distillation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 10659-10668.

[34] YU M Z, TAN S H, WU K L, et al. CORSD: Class-ori-

ented relational self distillation[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

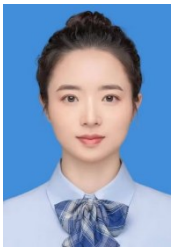
[35] ABNAR S, ZUIDEMA W. Quantifying attention flow in transformers[EB/OL]. (2020-05-02)[2024-12-16]. https://arxiv.org/abs/2005.00928.

作者简介

卿宇寒 男,2001年1月生,四川成都人. 重庆邮电大学通信与信息工程学院研究生. 主要研究方向为双模行为识别、计算机视觉和机器学习.
E-mail: qing_yh@foxmail.com

谭卓林 女,1998年1月生,四川达州人. 重庆邮电大学博士. 主要研究方向为视频分析、图像处理和计算机视觉.
E-mail: tanzhuolin98@gmail.com

高陈强 男,1981年8月生,重庆人. 中山大学·深圳教授、博士生导师. 主要研究方向为图像处理、视频分析和机器学习.
E-mail: gaochq6@mail.sysu.edu.cn



刘芳岑 女,1995年6月出,重庆人. 重庆邮电大学博士. 主要研究方向为红外小目标检测.
E-mail: liufc67@gmail.com