时空关键区域增强的小样本异常行为识别

肖进胜[1)] 王澍瑞[1)] 吴原顼[1)] 赵持恒[1)] 陈云华[2)] 章红平[3)]

[1)](武汉大学电子信息学院 武汉 430072)

[2)](广东工业大学计算机学院 广州 510006)

[3)](武汉大学 GNSS 研究中心 武汉 430072)

摘 要 异常行为识别在维护社会安全稳定方面起着重要的作用,相比于常见的正常行为识别,它是一项更具挑战性的任务。其难点主要体现在:异常行为实际发生的概率较低,因此可用于训练的样本数目相对较少;监控视频中,包含判断性信息的异常行为特征往往只存在于局部的关键区域中;异常行为时空变化复杂,导致连续地定位并利用关键区域特征变得更加困难。为了解决上述难题,本文提出时空关键区域增强的小样本异常行为识别方法,通过学习大规模正常行为数据集中的共性知识实现对数量较少的异常行为的识别,并选取视频中的关键区域对异常行为特征进行增强。特征向量由于其中的信息被压缩,而难以准确地定位关键区域,本文创新性地挖掘特征图中的二维空间信息,以自适应地选取异常行为的关键区域。单个的视频帧很难反映行为的变化情况,因此需要根据时空信息动态地选取关键区域。本文提出在特征图级别将长时间范围内的时序信息和短时间范围内的运动信息进行关联,以使关键区域有效地捕捉异常行为的连续变化。最后提出时空精细化小样本损失函数,以保证模型有效学习到在时间和空间中更精细化等级的特征。本文在 HMDB51、Kinetics 以及 UCF Crime v2 数据集上进行了实验,结果证明本文方法识别效果优于其他方法,在异常行为数据集上相对于最强的竞争者准确率提升了 0.6%。

关键词 异常行为识别;小样本学习;关键区域增强;时空精细化损失;时空关联

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2025.00068

Anomalous Action Recognition with Spatio-Temporal Key Region Enhancement and Few-Shot Learning

XIAO Jin-Sheng[1)] WANG Shu-Rui[1)] WU Yuan-Xu[1)] ZHAO Chi-Heng[1)]

CHEN Yun-Hua[2)] ZHANG Hong-Ping[3)]

[1)](School of Electronic Information, Wuhan University, Wuhan 430072)

[2)](School of Computer, Guangdong University of Technology, Guangzhou 510006)

[3)](GNSS Research Center, Wuhan University, Wuhan 430072)

Abstract Anomalous action recognition plays an important role in early-warning systems and has significant application merit for maintaining public security. The manual methods can lead to inefficiencies because of the tiredness, so we seek an automatic and smart method. Anomalous action recognition is a more challenging task compared to common normal action recognition, which is mainly reflected in the following points: few anomalous action videos can be collected due to the small probability of anomalous action occurring; because anomalous actions are often captured by surveillance cameras, the informative object features only exist in local key areas; there are complex spatio-temporal changes in the video, which increases the difficulty of locating and utilizing the features of key regions continuously. Based on the above analysis, we propose the anomalous action recognition method with spatio-temporal key region enhancement and

收稿日期:2023-12-23;在线发布日期:2024-09-10。本课题得到国家重点研发计划(2021YFB2501104)、湖北省重大攻关项目(尖刀2023BAA026)资助。**肖进胜**(通信作者),博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为视频图像处理、计算机视觉。E-mail: xiaojs@whu.edu.cn。**王澍瑞**,硕士研究生,主要研究方向为视觉三维感知。**吴原顼**,硕士,主要研究方向为视频分析。**赵持恒**,硕士,主要研究方向为图像处理。**陈云华**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、模式识别。**章红平**,博士,教授,主要研究领域为卫星惯性导航与组合导航。

few-shot learning in this paper. With the help of meta-learning, the network recognizes the limited anomalous action data by learning the task-level knowledge from the large normal action data. Because the global feature cannot highlight the key region information, the net also selects the key region to make feature enhancement. Similar methods select the key region by the feature vector, which has already lost the spatial information because of the spatial pooling. Our network selects the key region adaptively using the spatial information from the feature map. Without increasing the model weight, the distributed information at different positions on the feature map is fused to locate the key region. The selected region can be in any location where discriminative information exists to deal with ever-changing situations. The spatio-temporal information plays an important role in precise action modeling, and only using the individual frame cannot provide enough information for key region selection. The existing networks for spatio-temporal modeling can be large and high-cost. To improve computational efficiency, our work uses the lightweight module to build the relations between feature maps. The long-range temporal and short-range motion information at feature map level are both related to make the comprehensive modeling of the video. After the relation, the key regions selected can capture the continuous dynamic changes of the anomalous action. Finally, the spatio-temporal refined loss is proposed to ensure the learning of the feature at the more fine-grained level. The loss is calculated by multi-level spatial and temporal features. It reinforces the model to automatically look for the class-specific features, improving the generalization of the model. Using naive ways to train the network leads to poor performance. Three training tricks are proposed to make one-stage effective end-to-end training, improving the performance and reducing the training cost. We conduct experiments on the HMDB51, Kinetics, and UCF Crime v2 datasets, and the results show that our method exhibits superiority to the state-of-the-art counterparts on the few-shot action recognition, especially anomalous action recognition. Without direct training on the anomalous action dataset, it is worth noting that our method achieves 1.5% absolute boosts. The ablation studies demonstrate the effectiveness of the proposed modules, implying the potentiality of our plug-and-play modules to help improve other networks.

Keywords　anomalous action recognition; few-shot learning; key region enhancement; spatio-temporal refined loss; spatio-temporal relation

1　引　言

公共场景中可能发生具有危害性的异常行为，如打斗、爆炸、纵火等，若不及时处理会对人身甚至社会安全造成较大威胁，因此需要在监控视频中识别这些异常行为[1]。受到人力长期工作中误检和漏检问题的影响，大量监控视频未能得到有效识别。本文尝试提出一种智能算法，以实现监控视频中异常行为的自动识别。

目前针对异常行为识别的研究存在许多难点：(1)实际异常行为相对于正常行为发生的概率较低，因而异常行为样本收集困难。许多异常行为数据集例如 UCSD[2] 和 ShanghaiTech[3] 等大部分包含的都是正常行为数据，仅在测试集中有异常行为数据[4]。UCF Crime[5] 和 UCF Crime v2[6] 数据集包含有多种异常行为，但相较于包含了大量正常行为的数据集，如 Kinetics[7]，这两个数据集的种类和样本的数量都存在明显劣势，如果直接用作训练集效果较差[5]；(2)相较于正常行为，异常行为往往具有更复杂多样的表现形式，这需要提取更精准的特征表示[8]。一些现有模型只提取了视频帧的全局特征[9-11]。然而在监控视频中具有异常行为信息的目标，例如人、危险器械等，由于离监控摄像头较远，画面占比较小，通常只存在于局部区域。这种局部区域对行为准确分类的贡献比一般区域大，在本文中被称为关键区域。只使用全局特征难以凸显关键区域；(3)由于行为中存在复杂且多变的时空变化，在异常行为中起关键作用的区域需要以时空信息作为依据确认[12-13]，例如在打斗视频中不断前后躲闪和

有激烈扭打动作的主体所在的区域应被判为关键区域,而周围静止的背景和围观的个体则不是。因此,仅利用单一帧的空间信息是不充足的,这也增加了对关键区域的定位和利用的挑战。

为了解决上述难题,本文分别进行了针对性的分析与设计。直接使用异常行为训练难以得到具有良好泛化性的模型,而一些异常检测方法利用对大量的正常行为特征的学习来得到有效的模型[14-16],可以看出正常行为能够提供充足的训练数据。考虑到异常行为与正常行为数据量之间的差距,异常行为可视作小样本行为种类,而小样本学习认为从大规模正常行为数据集上学习到的与具体的行为种类无关的共性,例如对大量复杂行为的特征的表示与区分能力,能够帮助对样本较少的异常行为种类的识别。因此,本文提出基于小样本学习的方法,利用数量更多的正常行为样本帮助模型对异常行为的识别,以解决异常行为样本数量不足的问题。

考虑到全局特征描述异常行为存在一定的局限性,而在局部往往存在对异常行为识别作用较大的关键区域,因此本文提出空间自适应选取关键区域,并提取关键区域局部特征以增强全局特征。由于引入目标检测等额外的网络会提高训练和计算成本,本文提出了轻量化的模块选取关键区域。一些关键区域选取方法选择使用全局特征向量作为输入[12-13],但特征向量作为深层级的特征表示,原先的二维信息被压缩为了一维。为了尽可能补充先验知识,本文的关键区域选取模块利用了特征图中的二维空间信息以准确定位关键区域。相较于同类方法,该模块也有效降低了训练成本,在不需要区域坐标标注以及不增加参数的前提下,能够实现有效的区域选取。另外,该模块还保证了网络的梯度连续,能够实现端到端的训练,选取出的关键区域坐标精确,能够适应视频中复杂变化的情况。

一些方法在利用局部信息时只考虑了单帧的信息[17],忽视了在时空上的变化。本文提出了长短时特征图时空关联模块,充分利用多元时空信息以动态地关注时空关键区域。时空建模从不同范围进行考虑[9,18]长时间范围内连续时序上的信息以及相邻帧之间的运动信息。而专门用于提取这两种信息的三维卷积网络和光流网络参数量巨大,不适合移植于现有网络。本文使用两个轻量的子模块对特征图分别进行时序和运动关联,通过合并两种特征图实现完整的时空关联,经过多帧关联的特征图能更关注异常行为发生的区域。

对特征进行增强后,还需要有效的损失函数以确保学习过后的特征具有泛化性。传统的小样本学习损失函数只针对网络最终的输出特征,针对的特征种类单一。本文提出在空间和时间两方面对小样本学习损失函数进一步精细化,以确保模型能够对网络中间输出的不同类型的特征,包括不同子空间以及不同时序等级的特征,注入有效的语义信息。

本文尝试为之后的异常行为识别研究提供一种可行的模型基线和训练数据,主要贡献如下:

(1)本文提出了一种即插即用的空间自适应关键区域选取模块,能以特征图作为输入并利用其中的二维空间信息自适应地获取关键区域坐标,进而从局部突出异常行为的特点。在基本没有增加模型参数量的前提下,保证了模块的梯度连续以及输出的坐标精确。

(2)本文提出了一种长短时特征图时空关联模块,该模块能轻量地关联二维卷积神经网络提取出的各帧特征图中存在的时空信息,包括长时间范围内的时序信息和短时间范围内的运动信息等,能够在具有复杂时空变化的异常行为中更准确地选取关键区域。

(3)本文提出了一种时空精细化损失函数,通过增强更多等级特征的可靠性以提升模型整体的鲁棒性。该损失函数在空间上符合增强后特征的特点,实现了多个子空间有效信息的联合学习。同时为了使模型能从多个时间等级区分不同行为,在时序上增加了更精细的帧级别特征进行匹配。

2 相关工作

目前针对异常行为的研究主要存在的问题包括:数据量匮乏、异常行为特征难以有效构建等。针对上述问题,在当前行为识别方法中,一些算法在提升识别效率上取得了成功,而另一些算法则在利用局部区域上取得了突破性进展。在小样本行为识别领域,对行为特征的有效构建是提升效果的重要因素,大量方法在提取行为特征时考虑了视频帧之间的时空信息和帧内部的局部信息等。因此,本文将基于行为识别和小样本行为识别的相关成果,研究对于异常行为识别的有效方法。

2.1 行为识别

针对视频中的行为进行识别是目前视频理解领域的重要研究内容[19]。当前工作为了兼顾识别的精度和效率,提出使用较轻量化的二维特征提取网络作为主干,在此基础上利用特殊结构提取视频中

不同范围和等级的信息。例如 ACTION-Net[9] 在二维卷积网络基础上分别进行了时空信息、通道信息和运动信息的激活。STM[18] 提出了通道级别的时空模块和运动模块，并基于这两种模块构建了新的二维主干网络。

另一些研究工作发现行为识别存在空间冗余的问题，因此提取出视频中最具有信息的区域以提升识别效率成为研究热点，这方面最具代表性的是 AdaFocus[12] 和 AdaFocus V2[13]。AdaFocus 网络是首批研究降低空间冗余性来提升视频行为识别效率的方法，但其训练过程较为复杂。AdaFocus V2 在此基础上进行了改进，实现了端到端的训练，并能够输出精准的关键区域坐标值。本文则充分利用已被提取出的二维全局特征图自适应地选择关键区域。

2.2　小样本行为识别

目前的小样本识别算法大多都基于度量学习[20-21]，即学习一种特征嵌入方式，使得同类的样本在特征空间聚集，这种方式的准确分类需要构建有效的行为特征，因此对视频中具有代表性的时空信息的利用是近来研究工作的重点。TRX[10] 提出组合视频帧构建高阶的时序特征，以匹配不同速度和不同时间位置的行为。STRM[17] 在 TRX 基础上，在帧的全局联系不同帧中的目标信息，在局部区域寻找行为的外观特征。从上述工作可看出，帧局部空间信息和不同帧之间的时空信息是小样本行为识别的两种关键信息。为了充分利用这两种信息，本文关联时空信息辅助选取视频中的关键区域，并且利用关键区域的局部特征对全局特征进行增强。

2.3　异常行为识别

将异常行为从正常行为中检测出来有重大现实意义。然而由于异常行为样本较少，用于训练的数据量不充足，近年来许多工作通过对数量更多的正常行为样本进行学习，以实现对样本数量少的异常行为的检测，例如 Park 等人[14] 利用记忆模块存储多种正常行为数据的原型模式，并利用特征紧密性和分离性损失来保证记忆模块中正常数据的多样性；Liu 等人[15] 提出多级记忆模块确保更好地记忆正常数据以灵敏地检测异常，同时利用重建后的光流图以及先前视频帧的混合来预测未来帧；Chang 等人[16] 设计了一种自编码器结构以捕捉正常行为的空间和时间信息；Xing 等人[22] 将正常行为的特征分别存储在多个单元中，以减小正常行为的重建误差并增大异常行为的重建误差；Aich 等人[23] 提出了一种创新性的常态分类器来学习正常行为与伪异常行为特征的差别，实现在不需要跨域转换的条件下对异常行为的检测。进一步地，在检测出异常行为的基础上对异常行为类别进行细分有助于更有效地维护社会安全。Sultani 等人[5] 提出了包含 13 种异常行为的公开数据集 UCF Crime，并且提出了两种异常行为识别基线方法，分别使用 C3D、TCNN 作为主干网络，Maqsood 等人[24] 也是基于三维卷积构建网络以提取异常行为特征。但是这些算法难以有效建模复杂的异常行为，在 UCF Crime 上效果不好。受限于有效异常行为的数据量，后续针对异常行为识别的工作也较少。

3　本文算法

3.1　网络框架

为了解决异常行为识别中存在的难题，本文提出了一种时空关键区域增强的小样本异常行为识别网络。在引入小样本学习实现在样本很少时（1 个或 5 个）异常行为识别的基础上，本文网络提出提取时空关键区域特征以进行增强。网络每次训练都进行 C-way K-shot 小样本学习任务，也就是学习如何根据 C 类行为的包含 K 个实例的支撑集的特征与行为标签，判断预测集中未知样本的行为种类，因此网络每次输入包含支撑集和预测集的视频样本，每个视频样本是从原视频中稀疏采样得到的视频帧集合。网络整体结构如图 1 所示。

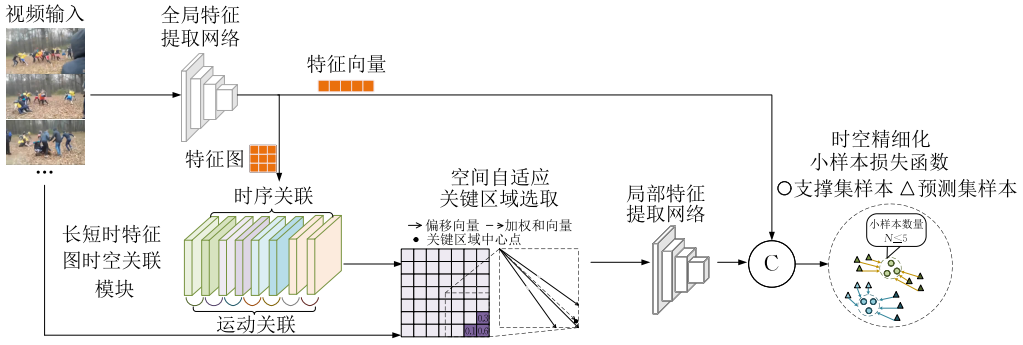


图 1　网络框架

从图 1 中可以看出在使用全局特征提取网络的基础上,本文网络增加了一条支路以定位并提取出有效的关键区域特征。这条支路通过关联丰富了各个特征图的时空信息,利用特征图上的二维空间信息自适应地得到有效局部区域。具体的流程为全局特征提取网络提取出每个视频帧的全局特征图和全局特征向量,将全局特征图输入局部信息表征支路。首先进入长短时特征图时空关联模块,关联后的特征图输入空间自适应关键区域选取模块,选取到的小尺寸关键区域输入局部特征提取网络得到局部特征向量,局部特征向量和全局特征向量连接得到最终代表视频样本的特征向量,最后支撑集和预测集的特征向量计算得到时空精细化小样本损失函数,以使不同空间来源和时序等级的特征学习到有效信息。

3.2 长短时特征图时空关联模块

在定位行为中的关键区域时,重要的依据不应是某一帧单独的空间信息,而应是相关目标在时空上产生的变化。而全局特征图中只存在单一帧的二维空间信息,因此提出对特征图进一步有效处理。具体的,运动信息反映出短时间范围内相邻帧之间的变化,能够有效捕捉当前帧中剧烈变化的区域;时序信息从长时间范围内连续堆叠的视频帧中获得,表征视频整体的时空变化趋势。这两种信息之间存在互补的关系,需要有效的方法综合利用。因此,基于较为轻量化的卷积模块[9,18],本文提出了长短时特征图时空关联模块,如图 2 所示,包含时序关联模块与运动关联模块,最后将两个模块输出的特征图相加以得到完整时空信息的特征图。本文的长短时特征图时空关联模块针对输入的全局特征图,在同一视频中关联了不同帧之间的时序信息和运动信息,最终对每一帧输出通道归一的二维空间掩码。

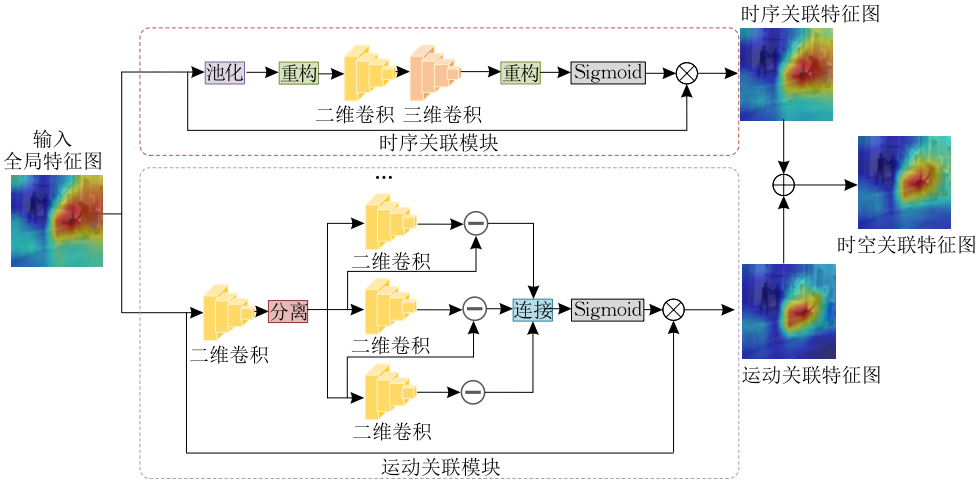


图 2 长短时特征图时空关联模块

3.2.1 时序关联模块

充分利用时序信息可以确保选取关键区域符合视频中长时间范围内的连续变化。引入具有时间维度和两个空间维度的三维卷积,可以对连续帧堆叠成的三维信息处理完成时序关联。三维卷积核 3×3×3 一次卷积只能混合连续三帧的空间信息,只使用三维卷积对于视频中连续的帧不能做到充分混合,而如果过度增加三维卷积核的尺寸,会带来大量的参数和训练成本。因此提出首先使用一个二维卷积混合所有视频帧信息,之后的三维卷积在不增加尺寸的条件下一次处理包含更多时序信息的特征,间接地扩大了三维卷积的关联范围。

具体过程如图 2 中“时序关联模块”所示,对于输入视频帧特征 $I\in \mathrm{R}^{N\times T\times C\times H\times W}$,其中 N 为批处理大小,T 对应时间维度上连续 T 帧特征,C 为通道数,H、W 对应特征高和宽。首先对特征进行通道平均归一池化,得到 $F\in \mathrm{R}^{N\times T\times 1\times H\times W}$,为了使之后二维卷积和三维卷积能够有效处理时间维度,对特征进行数据重构得到 $F'\in \mathrm{R}^{N\times 1\times T\times H\times W}$,接着通过输入与输出通道数都为 T 的 1×1 的二维卷积,如式(1)所示。

$$F''_t=\sum_c K^c F'_{t+c} \tag{1}$$

其中 K^c 是第 c 个通道的二维卷积参数,t 代表时间维度,F''_t 是二维卷积输出的第 t 帧特征。由于不使用分组卷积,卷积每个通道的输出特征都充分混合了来自所有输入通道的特征,此时每个输入通道对应一个时间点上的视频帧,因此能够以 1×1 的卷积大小对不同帧相同的空间位置信息进行混合。之后通过单通道输入输出的三维卷积对时序信息进行充分关联,如式(2)所示。

$$F'''=\sum_{x,y,z}K_{x,y,z}F''_{t+x,w+y,h+z} \tag{2}$$

其中 $K_{x,y,z}$ 是第 x 帧第 y 列第 z 行的三维卷积参数，t,w,h 代表时间维度和两个空间维度，F''' 为三维卷积输出的特征。再进行重构恢复维度顺序得到 $F^*\in\mathrm{R}^{N\times T\times 1\times H\times W}$，经过 Sigmoid 激活函数后与输入全局特征图 I 相乘得到时序关联特征图。

3.2.2　运动关联模块

运动信息表征相邻帧之间存在的时空变化，是判断关键区域的重要依据。例如对于行为的主体，包括人和车辆，当前运动的方向和速度影响了关键区域移动的方向和速度。为了有效建模运动信息，一种经典的方法是提取光流图进行表征，然而使用光流网络会额外引入大量参数和训练成本，同时降低了时间效率。本文利用轻量的模块，在特征图级别对运动信息进行关联，具体为基于帧差法构建运动关联模块，通过计算帧与帧特征之间的差值关联运动信息。

具体过程如图 2 中“运动关联模块”所示，对于输入 $I\in\mathrm{R}^{N\times T\times C\times H\times W}$，由于本模块针对的是特征图级别的信息的关联，首先通过一个 1×1 二维卷积对通道进行归一化 $M\in\mathrm{R}^{N\times T\times 1\times H\times W}$。接着为了便于计算不同帧之间的变化量，在时间维度上将各个帧的特征进行分离，得到 $M'\in\mathrm{R}^{N\times C\times H\times W}$。之后将各帧的特征通过 3×3 二维卷积再次进行空间编码，最后将编码后的特征与前一帧未经过编码的通道归一化特征作差，如式(3)所示。

$$M''_{t+1}=\sum_{i,j}K'_{i,j}M'_{t+1,w+i,h+j}-M'_t \tag{3}$$

其中 $K'_{i,j}$ 是第 i 列第 j 行的二维卷积参数，M''_{t+1} 是输出的第 $t+1$ 帧运动信息。将式(3)得到的帧之间的差值进行拼接，并在原来最后一帧的位置补零，经过 Sigmoid 激活函数后与输入全局特征图 I 相乘输出运动关联特征图。

3.3　空间自适应关键区域选取

卷积神经网络提取出的特征图上的元素与原图像的像素是存在空间对应关系的，因此特征图本身就具有一定的空间信息。利用特征图自身的空间信息在原图上提取一些感兴趣区域在目标检测领域已被证明可行，例如 Faster R-CNN[25] 和 Mask R-CNN[26] 中提出或改进的 RPN 方法。AdaFocus[12] 或 AdaFocus V2[13] 将特征图池化为特征向量并选取关键区域，没有充分利用特征图中的空间信息。本文提出以全局特征图作为输入自适应选取出关键区域，同时保持了网络梯度的连续，并且没有额外增加参数，输出关键区域坐标值精确到小数。

在固定了选取出的关键区域的尺寸和形状后，只需要确定关键区域上的一点就可以确定整个关键区域，因此空间自适应关键区域选取模块输出的是关键区域的中心点。空间自适应关键区域选取流程如图 3 所示。

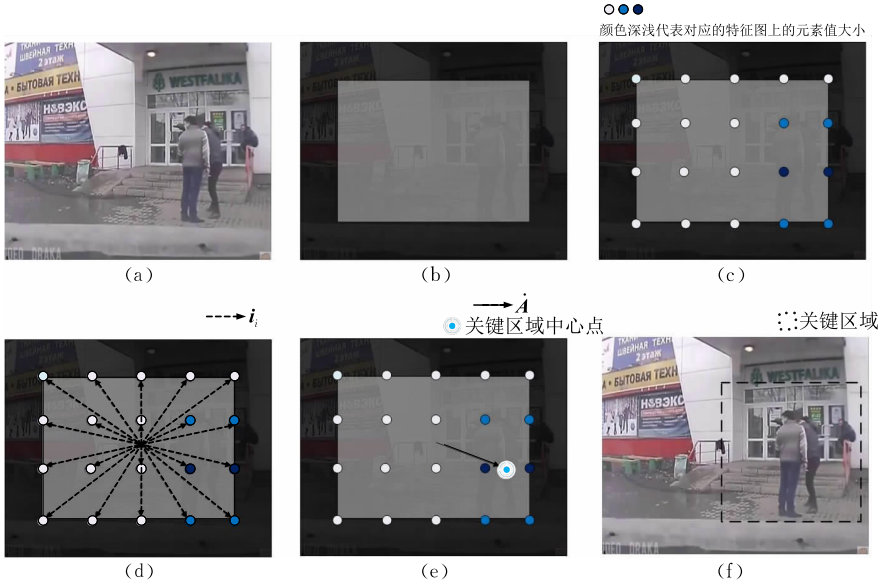


图 3　空间自适应关键区域选取流程

对于输入的视频帧如图 3(a)所示，首先根据选取关键区域尺寸和形状确定关键区域中心点的选取范围，为了确保选定中心点后，原图中有足够的空间能够提取关键区域，关键区域的选取范围与原图的边缘之间有一定距离，因此中心点 (x_c,y_c) 选取范围为

$$\begin{cases} l_a\le x_c\le h_a \\ l_a\le y_c\le h_a \end{cases} \tag{4}$$

可选范围如图 3(b)中浅色区域所示。为了充分利用特征图的空间信息,在该区域按照特征图的大小均匀地取 $N\times N$,共 L 个点,第 k 行第 j 列的点 (x_j,y_k) 的具体位置可被表示为

$$\begin{cases} x_j=l_a+(j-1)/(N-1)\times(h_a-l_a) \\ y_k=l_a+(k-1)/(N-1)\times(h_a-l_a) \end{cases} \tag{5}$$

每个点都按空间位置关系对应全局特征图上相应的元素,如图 3(c)所示(以 5×4 特征图为例)。为了进一步建立选取的 L 个点与特征图元素的关系,定义中心区域的中心点指向各个点的向量为“偏移”向量 $\boldsymbol{i}_i$,显然 $\boldsymbol{i}_i$ 与各个点和中心点的距离和方向有关,认为特征图上各元素 u_i 学习到的是对应点“偏移”向量的权重,如图 3(d)所示。最后就可以将选取的 L 个点根据特征图空间信息进行融合,此时将各个向量进行加权求和后得到和向量 $\dot{\boldsymbol{A}}$,如式(6)所示。

$$\dot{\boldsymbol{A}}=\sum_{i=1}^{L}u_i\boldsymbol{i}_i \tag{6}$$

为了使中心点最终落在图 3(b)中有效的范围内,对和向量 $\dot{\boldsymbol{A}}$ 指向的点 (x_t,y_t) 进行范围限定 $(g(x_t),g(y_t))$,最终得到关键区域的中心点,如图 3(e)所示,函数 g 如图 4 与式(7)所示:

$$g(i)=\begin{cases} h_a, & i\geq h_a \\ i, & l_a<i<h_a \\ l_a, & i\leq l_a \end{cases} \tag{7}$$

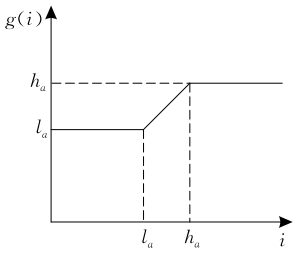


图 4 函数 g 示意图

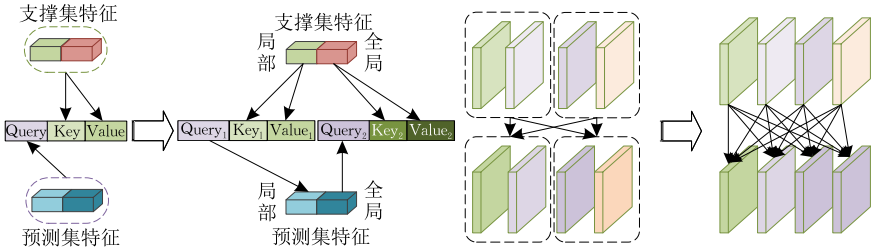
通过平移中心点坐标可以得到关键区域其他点的坐标。由于得到的中心点坐标值是在原图有效范围内带小数,得到的关键区域上的点与原图上的整数值像素点并不总是完全对应。保留带小数的坐标能够更精准地反映出关键区域的位置信息,也能更好地在连续变化的行为中选取关键区域。本文采用双线性插值的方法建立两个坐标之间的关系。与带小数值坐标 (x_{ij},y_{ij}) 最近的四个整数值坐标为 $(\lfloor x_{ij}\rfloor,\lfloor y_{ij}\rfloor)$,$(\lfloor x_{ij}\rfloor+1,\lfloor y_{ij}\rfloor)$,$(\lfloor x_{ij}\rfloor,\lfloor y_{ij}\rfloor+1)$,$(\lfloor x_{ij}\rfloor+1,\lfloor y_{ij}\rfloor+1)$,其中 $\lfloor\ \rfloor$ 代表向下取整,这四个坐标相应的像素值分别为 $(s_{ij})_{00}$,$(s_{ij})_{01}$,$(s_{ij})_{10}$,$(s_{ij})_{11}$,根据双线性插值可以得到连续坐标的像素值 s'_{ij},如式(8)所示:

$$\begin{aligned} s'_{ij}=&(s_{ij})_{00}(\lfloor x_{ij}\rfloor-x_{ij}+1)(\lfloor y_{ij}\rfloor-y_{ij}+1)+\\ &(s_{ij})_{01}(x_{ij}-\lfloor x_{ij}\rfloor)(\lfloor y_{ij}\rfloor-y_{ij}+1)+\\ &(s_{ij})_{10}(\lfloor x_{ij}\rfloor-x_{ij}+1)(y_{ij}-\lfloor y_{ij}\rfloor)+\\ &(s_{ij})_{11}(x_{ij}-\lfloor x_{ij}\rfloor)(y_{ij}-\lfloor y_{ij}\rfloor) \end{aligned} \tag{8}$$

最终在原输入视频帧通过裁剪得到关键区域,如图 3(f)所示。由于充分利用了特征图中的空间信息,选取出的关键区域能够包含对行为识别贡献较大的信息。并且在选取过程中利用的是已被提取出的特征图,基本没有额外增加模型的参数,同时能够保证梯度连续以进行端到端的训练,能起到即插即用的效果。得到的关键区域输入到局部特征提取网络中进行特征提取,提取出的局部特征与全局特征进行连接构成最终特征向量。

3.4 时空精细化小样本损失函数

获得增强特征后,还需要有效的损失函数使模型具有与具体类别无关的行为识别能力。TRM (Temporal-Relational Cross Transformer Module,时间关系交叉变换器模块)[10] 基于交叉注意力机制匹配未知行为,然而,TRM 主要针对的是输出的全局特征,缺乏对多空间特征的联合处理机制;并且构建的时序元组没有直接使用帧级的特征,可能使最终学习到的帧级特征缺乏有效性。为了解决上述问题,我们利用时间和空间上更精细的特征进行匹配,提出时空精细化小样本损失函数,如图 5 所示。它由空间精细化匹配损失和时序精细化匹配损失两部分的和组成。



(a)空间精细化匹配 (b)时序精细化匹配

图 5 时空精细化小样本损失函数示意图

3.4.1　空间精细化匹配损失

TRM[10]中的每个交叉注意力转换器只使用了一组键值对。而在本文中，除了全局特征外，最终特征还包含用于增强的局部特征。考虑到全局特征和局部特征表征不同语义的信息，使用单一注意力机制不能体现两个子空间的区别。因此对于最终特征的小样本损失函数计算，引入多头注意力机制。对全局和局部的子空间使用不同的编码器，以对不同类别的信息分别进行具体的映射。

具体地，对于来自支撑集的样本，将其增强后的特征 F_{se}，分为全局部分 F_{sg} 和局部部分 F_{sl}。类似的对于来自预测集的样本，其增强后的特征 F_{qe} 分为 F_{qg} 和 F_{ql}，并且分别使用不同组的键值对计算注意力权重，如式(9)所示：

$$\begin{cases} S_l=\mathrm{Softmax}\left(\dfrac{F_{ql}K_{sl}^T}{\sqrt{d_{sl}}}\right)V_{sl} \\ S_g=\mathrm{Softmax}\left(\dfrac{F_{qg}K_{sg}^T}{\sqrt{d_{sg}}}\right)V_{sg} \end{cases} \tag{9}$$

其中 K_{sl}、V_{sl} 和 K_{sg}、V_{sg} 分别由 F_{sl} 和 F_{sg} 通过线性转换得到，d_{sl} 和 d_{sg} 分别为 K_{sl} 和 K_{sg} 的维数。基于多头注意力机制将两部分聚合，并计算得到不同样本之间的距离。

$$D_{MH}=\mathrm{concat}(S_g,S_l)-F_{qe} \tag{10}$$

根据上式即可求出支撑集和预测集中样本之间的距离并匹配样本的行为种类。由于在损失函数中加入了多头注意力机制，模型能够自动地注意并表征不同来源的信息。空间精细化匹配损失与之前提出的关键区域增强能够相互促进，保证特征表征力和行为识别准确度的同步增强。

3.4.2　时序精细化匹配损失

通过对两帧或三帧进行组合在时序上构建高等级的特征能够有效提升表征力，在此基础上强化对低等级特征的学习能够使模型从不同方面学习如何区分不同的行为，进而防止出现过拟合以提升模型的泛化性。本文引入了一种使用全范围时序信息进行准确匹配的损失函数。对于复杂的异常行为，可以进行精细化分解，如打斗，往往包含拳打、脚踢等一些子动作，但是对于不同视频中的行为，这些子动作往往存在于不同的位置，使用简单的相同时间位置匹配存在很大局限，因此需要灵活的匹配方式来解决潜在的不对齐问题。

将行为看作是不同帧级别的子动作的集合，为了精细化地匹配不同视频中的子动作，使用双向平均豪斯多夫距离[27]。具体地，对于来自支撑集和预测集的行为，分别提取包含帧级别的特征集合为 $F_s=\{f_s^0,f_s^1,\cdots,f_s^i\}$ 和 $F_q=\{f_q^0,f_q^1,\cdots,f_q^j\}$。两个集合之间的时序精细匹配距离为

$$D_{TR}=\frac{1}{d}\sum_{f_s^i\in F_s}\left(\min_{f_q^j\in F_q}\|f_s^i,f_q^j\|\right)+\frac{1}{d}\sum_{f_q^j\in F_q}\left(\min_{f_s^i\in F_s}\|f_q^j,f_s^i\|\right) \tag{11}$$

其中 $\|\ \|$ 代表帧级别特征之间的余弦距离，d 代表帧级别特征的维数。时序精细化匹配损失函数在使用全局时序信息的基础上，利用更细致等级的特征进行行为匹配。并且与之前的空间精细化匹配损失形成了互补，对子序列和帧等级的特征都进行了损失函数的计算，保证了模型在不同等级上特征的有效性。

3.5　训练技巧

由于网络中包含有两个特征提取网络：全局特征提取网络、局部特征提取网络，且这两个特征提取网络在功能上存在明显差异。为了保证不同的特征提取网络训练得到合适的功能，使用分阶段训练[12]容易导致训练成本大幅度增加。而使用一次简单的端到端的训练就得到具有有效泛化性的网络也存在很大困难。为了保证训练得到有效的参数[13]，本文加入了三种特殊的训练技巧：梯度提前中断、局部区域数据增强和额外的损失函数。

由于本文提出的模块梯度都是连续的，而整体网络是通过全局特征提取网络提取出的特征图得到关键区域，并从关键区域提取局部特征向量，因此在训练过程中，提取的局部特征向量通过反向传播也会更新全局特征提取网络的参数。考虑到全局特征提取网络的主要功能是提取对异常行为具有表征力的全局特征，局部特征向量的梯度输入到全局特征提取网络中更新会对此产生干扰，使得全局特征提取网络训练效果不理想。因此本文采用梯度提前中断的方式，在全局特征图输入之后模块前将梯度进行截断，使得局部特征提取网络与全局特征提取网络和长短时特征图时空关联模块梯度更新分离，以减少训练过程中各个模块之间的干扰，提升训练效果。

特征提取网络在训练时常常随机截取原图中的一部分，将其作为新的训练数据进行数据增强。而本文的局部特征提取网络在训练时，输入局部图像是由全局特征图唯一确定的，这使得训练数据相对单调，不利于提升模型泛化性。为了提升数据多样性，本文提出使用类似随机截取的数据增强的方式，

具体为在训练过程中有 50%的概率向关键区域选取模块输入完全随机产生的特征图向量,以得到随机的局部区域输入到局部特征提取网络中,提升局部特征提取网络的泛化性。

分阶段训练的方式通过预训练主干网络,确保其具有强大的特征提取能力,以提升最终总模型的效果。同样的,本文提出的网络中也应确保提升两个主干网络提取出的全局特征和局部特征的表征力。对全局特征的提升能够帮助对关键区域的准确选取,同时全局特征和局部特征的提升也直接帮助最终特征的提升。因此除了根据最终特征计算出的时空精细化小样本损失 $STR(F_{\text{final}})$外,对全局特征 F_{global}和局部特征 F_{local}也分别计算损失,由于本文基于 TRM 计算得到小样本损失函数,因此对全局特征和局部特征分别增加额外的 TRM 模块,计算得到相应的全局和局部特征小样本损失,最后与最终特征损失相加求平均得到总损失,如式(12)所示。

$$loss=\frac{1}{3}(STR(F_{\text{final}})+TRM(F_{\text{global}})+TRM(F_{\text{local}})) \tag{12}$$

4 实 验

4.1 实验设置

4.1.1 数据集

为了验证本文算法的有效性,本文在多种不同的数据集上进行了实验。

(1) HMDB51[28]数据集包含 51 种行为,该数据集中包含许多生活中常见行为,例如走路、奔跑、坐下等,也包含一些如持枪射击、踢、用拳猛击等通常被认为异常的行为。按照文献[29]标准将其中 31 种行为用于训练集,10 种行为用于验证集,10 种行为用于测试集。

(2) Kinetics[7]数据集包含 400 种行为和 306 245 种视频片段。按照 CMN-J[30]中提出的方式对 Kinetics 数据集进行划分,选取 100 种行为种类,每类包含 100 个视频,其中 64 类用于训练集,12 类用于验证集,24 类用于测试集。采用 Kinetics 数据集的训练集进行训练,并分别在 Kinetics 数据集的测试集和 UCF Crime v2 数据集上进行测试。

(3) UCF Crime v2 数据集,除去正常行为视频数据,UCF Crime 数据集包含 13 种异常行为和 950 个视频。在此基础上 UCF Crime v2,不仅在 UCF Crime 数据集的基础上增加了 2 种异常行为和 233 个视频,还对原训练集进行了时序标注。考虑到小样本行为识别任务中训练集和测试集行为种类不能有重叠,而 UCF Crime v2 数据集具有的异常行为种类相对较少,因此利用在行为种类和视频数量较多的 Kinetics 数据集上训练得到的模型在 UCF Crime v2 数据集上测试。

4.1.2 实验细节

本文实验采用了与之前方法相同的预处理过程。在模型训练时,从数据集中随机抽取样本分别组成支撑集和预测集,对每个视频稀疏采样出 8 个视频帧,这些视频帧高被调整为 256 并经过数据增强,包括随机水平翻转和裁剪(裁剪尺寸为 224×224)。测试也基本采取与训练相同的过程,在随机抽取样本组成支撑集和预测集之后,对视频的预处理过程与训练时基本相同,但是不进行数据增强,只对视频帧进行中心裁剪。网络特征提取后计算支撑集和预测集样本特征之间的距离,以判断预测集中样本的标签。本文网络使用 2×GeForce RTX 3090 Ti 进行训练,考虑到显存限制,设定训练时每次小样本学习任务中每一类行为预测集包含的样本数为 3。

本文网络使用 TRX[10]作为基线,对于提取的关键区域,设定关键区域尺寸为 128×128,全局特征提取和局部特征提取网络都使用在 ImageNet 上预训练的 Resnet50 初始化。采用随机梯度下降法训练模型,初始学习率为 0.001,对于 HMDB51 数据集训练迭代 20 000 次,对于 Kinetics 数据集训练迭代 30 000 次。测试时随机进行 10 000 轮,每一轮都统计异常行为识别的准确度(正确识别样本个数/总样本个数),最后报告平均精确度。

4.2 对比实验结果

在 HMDB51 和 Kinetics 等正常行为数据集以及 UCF Crime v2 异常行为数据集下与当前的 SOTA 方法进行对比实验。因为涉及异常行为数据集的结果,而一般的小样本行为识别方法没有在该数据集的结果,因此本文重新实现了多种经典算法。同时为了更公平地比较,也在相同的实验平台和设置下得到了在正常行为数据集下的结果,在对比时只考虑相同实验条件下的结果,而原论文中汇报的结果则只是作为参考。网络训练采取 C-way K-shot 小样本学习任务,基于 5-way 1-shot 和 5-way 5-shot 两种设置,也就是根据一共 5 类行为,每类行为有 1 个或 5 个的已知样本组成的支撑集,判断预测集中未知样本的种类,对比实验的结果如表 1 所示。

表 1　对比实验结果　　（单位：%）

方法	来源	HMDB		Kinetics		UCF Crime v2	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ARN[29]*	ECCV2020	45.5	60.6	63.7	82.4	—	—
OTAM[31]*	CVPR2020	—	—	73.0	85.8	—	—
OTAM[31]		49.6	60.5	68.1	79.4	**39.3**	47.9
TRX[10]*	CVPR 2021	—	75.6	63.6	85.9	—	—
TRX[10]		51.2	73.3	63.7	84.9	34.4	48.2
ATA[11]*	ECCV 2022	59.6	76.9	74.3	87.4	—	—
ATA[11]		**56.5**	71.5	**71.1**	85.0	39.0	48.4
STRM[17]*	CVPR 2022	—	77.3	—	86.7	—	—
STRM[17]		50.4	73.3	66.8	85.1	36.5	48.7
SloshNet[32]*	AAAI 2023	—	77.5	—	87.0	—	—
SloshNet[32]		50.9	74.0	63.0	85.8	36.8	49.3
BiMACL[33]*	ICASSP 2024	57.0	78.4	68.1	87.6	—	—
BiMACL[33]		53.3	73.7	64.8	85.6	36.6	49.6
本文方法		54.5	**74.3**	67.9	**86.1**	37.0	**50.2**

注：在对比本文方法时只考虑在同一平台和设置下的结果，“*”表示原论文中汇报的结果仅作为参考，不作为对比对象；“—”表示原文中结果空缺；加粗代表在该数据集和设置下最好的结果（其他表同样含义）。

表 1 中结果显示，当使用 5-way 1-shot 设置时，ATA 和 OTAM 的方法效果要更好，但是本文方法能够实现利用视频中的子序列判断行为，以排除可能产生干扰的信息，而不需要像 OTAM 和 ATA 使用整个视频进行匹配。同时本文方法在支撑集样本数量较多的情况下，例如 5-shot 设置下，相较于这两种方法也有明显的优势，表 1 实验结果也证明了这一点。根据实验结果对比，本文方法相对于基线方法 TRX、STRM、SloshNet、BiMACL 等使用相同基线的方法以及其他方法如 ARN 等，在 1-shot 设置下效果具有明显提升，验证了方法的有效性。从上述分析和实验结果对比[10-11,25]也可以看出，本文方法与 ATA 和 OTAM 两种方法的效果差距主要原因在于使用的基线方法有不同的特点，本文使用的基线在 1-shot 设置下效果较差，但是能构建更丰富的高阶时序特征实现更灵活的匹配，因此当样本数量变多时，能够在复杂多样的行为时空特征中精准匹配而显现出优势。

在 5-shot 设置下，本文方法在正常行为数据集上取得了最好的效果。这表明利用特征图关联后自适应选取的关键区域能够包含有效的语义信息，并实现了对正常行为特征的有效增强。这也展示了我们的方法对于大部分视频行为分类任务的潜能。对于更具有挑战性的异常行为识别，本文方法有较大的提升效果，相对于最强的竞争者准确率提升了 0.6%。这表明提取的关键区域能够有效解决异常行为识别的难题，在整个视频区域中，凸显了对异常行为识别的关键性信息以及抑制了可能产生干扰的背景信息。图 6 是在异常行为数据集 UCF Crime v2 下的一些可视化实验结果，从上而下分别展示集中打斗、横幅、抢劫、纵火、扔燃烧瓶等异常行为的示例，并且使用红框标出了选取的时空关键区域。这些示例能够在视觉层面辅助证明自适应提取出的区域能够包含异常行为主体（具体的人）以及火光横幅等明显特征，同时由于在特征图等级关联了时空信息，在异常行为连续变化的过程中能够动态地定位出关键区域。

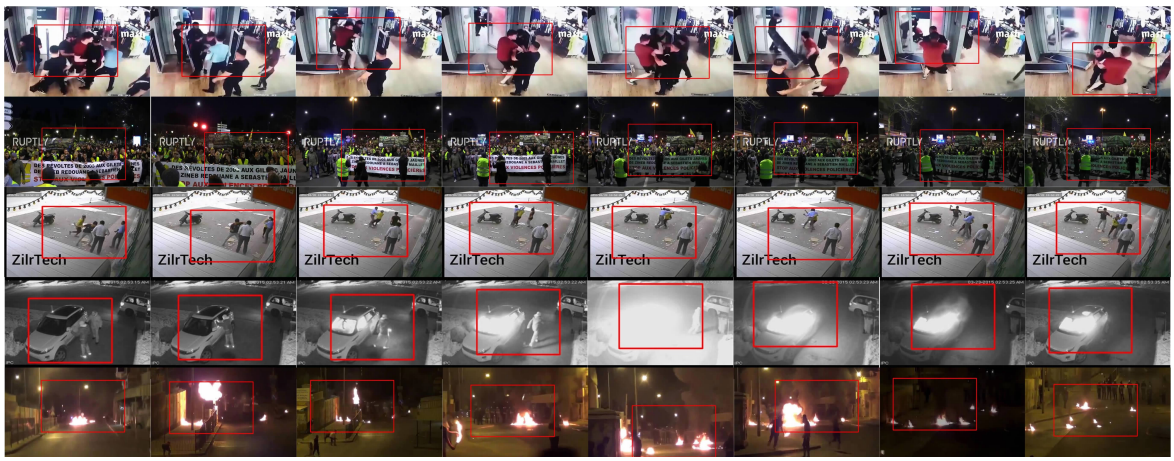


图 6　关键区域选取可视化结果

4.3 消融实验结果

为验证提出模块的有效性，在三个数据集上使用 5-way 1-shot 和 5-shot 进行了消融实验。实验过程中在基线方法的基础上，逐渐加入提出的模块，结果如表 2 所示。首先加入空间自适应关键区域选取，结果证明在没有显式增加参数的前提下，自适应选取出的关键区域能够凸显在空间上更具有代表性的信息，并且提取的局部区域特征能够有效增强行为特征。长短时特征图时空关联模块引入了视频中的时空信息，使得选取的关键区域包含了更多有效的深层语义特征，提升了识别的准确率。最后加入的时空精细化小样本损失函数使模型在多个等级上自动地学习到能够区别不同类的特征。

表 2 消融实验结果 （单位：%）

基线	空间自适应关键区域选取	长短时特征图时空关联	时空精细化小样本损失函数	HMDB		Kinetics		UCF Crime v2	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
✓				51.2	73.3	63.7	84.9	34.4	48.2
✓	✓			52.8	73.7	66.8	86.0	36.1	49.5
✓	✓	✓		53.9	74.0	67.5	86.1	36.9	50.0
✓	✓	✓	✓	**54.5**	**74.3**	**67.9**	**86.1**	**37.0**	**50.2**

为了更直观地体现长短时特征图时空关联模块的效果，本文也分别对时空关联前后的特征图在 UCF Crime v2 数据集上进行了可视化，如图 7 所示，其中第一行是输入的原图，第二行、第三行分别是全局特征提取网络提取的特征图和时空关联后的特征图上采样之后叠加在原图上的结果，特征图上的部分红色越深代表受关注的程度越高，而蓝色越深代表受关注的程度越低。在关联之前，全局特征图由于只是根据单帧的空间信息得到，有时会关注一些静止的区域，而这些静止的区域并没有直接参与到异常行为中；而关联后的特征图由于全面利用了长时间和短时间范围内的时空信息，能够抑制对静止的背景等的关注，同时强调突出在整个视频和当前时间点下变化更激烈的区域，这也是异常行为主要发生的区域，因此更能准确地定位关键区域。

第 3.4 节提出了三种训练技巧，为了验证这三种技巧的有效性，本文在三个数据集上以 5-way 5-shot 设置进行实验。为了突出训练技巧的效果以及排除其他干扰，实验聚焦于模型参数量最大的部分，具体组成为在基线模型上增加了空间自适应关键区域选取以提取局部特征。在实验过程中逐渐增加 3.4 节中提出的技巧，结果如表 3 所示。

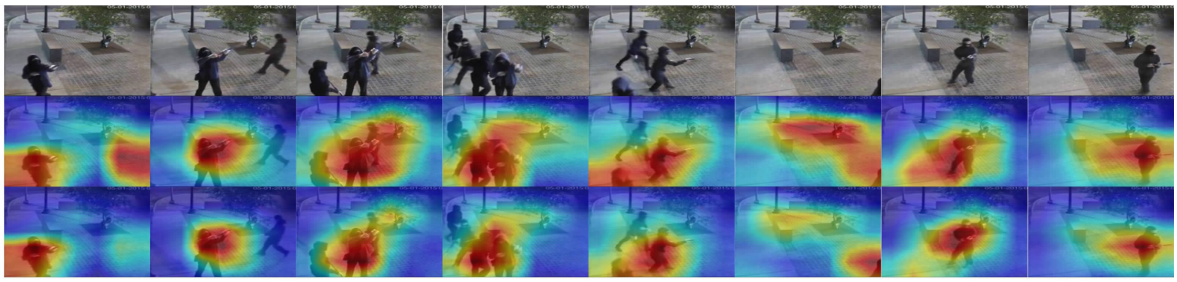


图 7 特征图可视化结果示例

表 3 训练技巧消融实验结果 （单位：%）

梯度提前中断	额外的损失函数	局部区域数据增强	HMDB	Kinetics	UCF Crime v2
			63.9	71.5	44.4
✓			72.3	84.5	47.5
✓	✓		73.3	85.5	49.3
✓	✓	✓	**73.7**	**86.0**	**49.5**

简单的训练使得模型参数学习的过程混乱，很难收敛到有效值。增加梯度中断技巧提升了训练的稳定度，保证了各个模块学习策略明确，提升了效果。增加额外的损失函数，提升了全局特征和局部特征的表征力，对行为建模效果更好。加入局部区域数据增强使得模型的泛化性进一步提升，全面提升了准确率。

本文还进行了实验来分析本文方法在 UCF Crime v2 数据集上产生误识别的原因：(1)在一些极端情况下，关键区域仍无法突出异常行为目标的

特征。这是由于本文为了轻量选取关键区域而固定了其尺寸，导致一些很小的目标在关键区域中仍然不显著，如图 8 所示。要解决这一问题可能需要引入针对细小目标的检测网络，但这也可能会增加模型的计算成本；(2) 选取的支撑集样本不合适。支撑集是用来判断未知样本种类的依据，对识别效果起到重要作用，本文为了体现模型效果采用随机抽取的方式构建支撑集，但当选取的支撑集自身存在模糊性时，会导致对未知样本的区分变得困难。例如图 9 所示，5 个样本都可以被概括为使用肢体或工具对其他人的袭击，然而却被分别用于代表 5 种不同的行为种类。要解决这一问题需要算法对支撑集样本进行过滤，或对数据集进行相关的工作。

图 8　关键区域失败案例

(a) 虐待　(b) 逮捕　(c) 袭击　(d) 打斗　(e) 抢劫

图 9　支撑集样本选取失败案例

5　结　论

本文提出了一种时空关键区域增强的小样本异常行为识别网络。基于小样本学习实现了利用大量正常样本数据中学习到的共性对异常行为种类的识别，解决了异常行为样本少的难题。针对异常行为的主体以及显著性特征占监控视频比例较小的问题，提出提取关键区域特征对全局特征进行增强。具体地，提出了一种利用特征图空间信息自适应地提取关键区域的方法，作为一种即插即用的模块能够在原图上选取坐标值精准的关键区域，同时基本没有额外增加模型大小。仅仅利用某一帧的空间信息没有充分挖掘出视频中的所有信息，为了有效利用长时间范围内的时序信息和短时间范围内的运动信息，提出长短时特征图时空关联模块，在特征图等级上将不同帧之间的信息进行关联，以提升异常行为时空动态变化过程中关键区域选取的有效性。网络最后计算时空精细化小样本损失函数以强化在多个等级上特征的有效学习。本文方法在多个数据集包括 HMDB51、Kinetics，特别是异常行为数据集 UCF Crime v2 上进行了对比实验和消融实验，实验结果证明了本文方法的有效性。

致　谢　本论文的数值计算得到了武汉大学超级计算中心的计算支持和帮助，在此表示诚挚的感谢！

参　考　文　献

[1] Xie Zhao, Zhou Yi, Wu Ke-Wei, et al. Activity recognition based on spatial-temporal attention LSTM. Chinese Journal of Computers, 2021, 44(2): 261-274(in Chinese)
(谢昭，周义，吴克伟等. 基于时空关注度 LSTM 的行为识别. 计算机学报，2021，44(2)：261-274)

[2] Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(1): 18-32

[3] Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked RNN framework//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 341-349

[4] Xiao Jin-Sheng, Guo Hao-Wen, Xie Hong-Gang, et al. Probabilistic memory auto-encoding network for abnormal behavior detection in surveillance videos. Journal of Software, 2023, 34(9): 4362-4377(in Chinese)
(肖进胜，郭浩文，谢红刚等. 监控视频异常行为检测的概率记忆自编码网络. 软件学报，2023，34(9)：4362-4377)

[5] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6479-6488

[6] Öztürk H İ, Can A B. ADNet: Temporal anomaly detection in surveillance videos//Proceedings of the International Conference on Pattern Recognition (ICPR) International Workshops and Challenges. Milan, Italy, 2021: 88-101

[7] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset//Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6299-6308

[8] Xiao Jin-Sheng, Shen Meng-Yao, Jiang Ming-Jun, et al. Abnormal behavior detection algorithm with video-bag attention mechanism in surveillance video. Acta Automatica Sinica, 2022, 48(12): 2951-2959(in Chinese)
(肖进胜, 申梦瑶, 江明俊等. 融合包注意力机制的监控视频异常行为检测. 自动化学报, 2022, 48(12): 2951-2959)

[9] Wang Z, She Q, Smolic A. ACTION-Net: Multipath excitation for action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 13214-13223

[10] Perrett T, Masullo A, Burghardt T, et al. Temporal-relational CrossTransformers for few-shot action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 475-484

[11] Nguyen K D, Tran Q H, Nguyen K, et al. Inductive and transductive few-shot video classification via appearance and temporal alignments//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 471-487

[12] Wang Y, Chen Z, Jiang H, et al. Adaptive focus for efficient video recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 16249-16258

[13] Wang Y, Yue Y, Lin Y, et al. AdaFocus V2: End-to-end training of spatial dynamic networks for video recognition//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 20030-20040

[14] Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 14372-14381

[15] Liu Z, Nie Y, Long C, et al. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 13588-13597

[16] Chang Y, Tu Z, Xie W, et al. Clustering driven deep autoencoder for video anomaly detection//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 329-345

[17] Thatipelli A, Narayan S, Khan S, et al. Spatio-temporal relation modeling for few-shot action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 19958-19967

[18] Jiang B, Wang M M, Gan W, et al. STM: Spatiotemporal and motion encoding for action recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 2000-2009

[19] Ding Chong-Yang, Liu Kai, Li Guang, et al. Spatio-temporal weighted posture motion features for human skeleton action recognition research. Chinese Journal of Computers, 2020, 43(1): 29-40(in Chinese)
(丁重阳, 刘凯, 李光等. 基于时空权重姿态运动特征的人体骨架行为识别研究. 计算机学报, 2020, 43(1): 29-40)

[20] Wang X, Zhang S, Qing Z, et al. MoLo: Motion-augmented long-short contrastive learning for few-shot action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18011-18021

[21] Liu S, Jiang M, Kong J. Multidimensional prototype refactor enhanced network for few-shot action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6955-6966

[22] Xing P, Li Z. Visual anomaly detection via partition memory bank module and error estimation. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3596-3607

[23] Aich A, Peng K C, Roy-Chowdhury A K. Cross-domain video anomaly detection without target domain adaptation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2023: 2579-2591

[24] Maqsood R, Bajwa U I, Saleem G, et al. Anomaly recognition from surveillance videos using 3D convolution neural network. Multimedia Tools and Applications, 2021, 80(12): 18693-18716

[25] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[26] He K, Gkioxari G, Dollár P, et al. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969

[27] Wang X, Zhang S, Qing Z, et al. Hybrid relation guided set matching for few-shot action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 19948-19957

[28] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition//Proceedings of the 2011 International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563

[29] Zhang H, Zhang L, Qi X, et al. Few-shot action recognition with permutation-invariant attention//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 525-542

[30] Zhu L, Yang Y. Label independent memory for semi-supervised few-shot video classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 273-285

[31] Cao K, Ji J, Cao Z, et al. Few-shot video classification via temporal alignment//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10618-10627

[32] Xing J, Wang M, Liu Y, et al. Revisiting the spatial and temporal modeling for few-shot action recognition//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(3): 3001-3009

[33] Guo H, Yu W, Yan Y, et al. Bi-directional motion attention with contrastive learning for few-shot action recognition//Proceedings of the ICASSP 2024－2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea, 2024: 5490-5494

XIAO Jin-Sheng, Ph. D., associate professor. His research interests include video and image processing, computer vision.

WANG Shu-Rui, M. S. candidate. His research interest is vision and 3D perception.

WU Yuan-Xu, M. S. His research interest is video analysis.

ZHAO Chi-Heng, M. S. His research interest is image processing.

CHEN Yun-Hua, Ph. D., associate professor. Her research interests include computer vision and pattern recognition.

ZHANG Hong-Ping, Ph. D., professor. His research interests include satellite inertial navigation and combined navigation.

Background

In public spaces, dangerous anomalous actions such as fighting, explosions, and arson can pose significant threats to personal and societal safety if not promptly identified and addressed. Video surveillance systems are crucial for recognizing these behaviors, but human monitoring can be prone to errors, leading to missed detections. To overcome these limitations, this study aims to develop an intelligent algorithm for automatically detecting anomalous actions in surveillance videos.

Several challenges exist in recognizing anomalous actions. Firstly, the occurrence of anomalous actions is rare compared to normal activities, making it difficult to collect sufficient anomalous samples. Most existing datasets, like UCSD and ShanghaiTech, predominantly contain normal behavior, with anomalous actions appearing only in test sets. Datasets like UCF Crime offer a variety of anomalous actions but still have significantly fewer samples compared to normal behavior datasets, leading to poor performance when used directly for training. Secondly, anomalous actions are often more complex and varied in nature, requiring precise feature extraction. Existing models generally focus on global features, which can overlook critical regions, such as distant individuals or dangerous objects, which are vital for accurate classification. Lastly, the spatio-temporal variability in behaviors complicates the identification of key regions, making it difficult to rely solely on spatial information from single frames.

To address these challenges, this paper introduces a method that leverages few-shot learning and spatio-temporal key region enhancement. Given the difficulty of training models with limited anomalous samples, the proposed method uses normal behavior data to help identify anomalies by treating them as a few-shot learning problem. The method employs a spatially adaptive key region selection module, which autonomously identifies and enhances critical local features within the feature map, improving the global feature representation. This module is lightweight and can be integrated into existing networks without adding significant computational overhead. Additionally, a spatio-temporal correlation module is introduced to capture the dynamic nature of anomalous actions by associating temporal and motion information across frames. This approach ensures that key regions are accurately identified even in complex scenarios. To further enhance the model's robustness, a refined spatio- temporal loss function is proposed, which injects semantic information into different feature levels, ensuring the model can effectively learn and generalize across various contexts. Experiments on the HMDB51, Kinetics, and UCF Crime v2 datasets demonstrate that the proposed method outperforms state-of-the-art approaches, particularly in few-shot and anomalous action recognition.

This work was supported by the National Key Research and Development Program of China (No. 2021YFB2501104) and the Major Program (JD) of Hubei Province (No. 2023BAA02604).