优化YOLO网络的人体异常行为检测方法

张红民[1,2],庄 旭[1],郑敬添[1],房晓冰[1]
1.重庆理工大学 电气与电子工程学院,重庆 400054
2.重庆理工大学 两江国际学院,重庆 401135

摘 要:鉴于公共场合监测视频信息中周围环境背景信息干扰大以及人体异常行为目标的尺度不同,目前人体异常行为检测的准确性难以进一步提高。针对上述问题,设计了通过改进YOLOv5网络的异常行为检测方法。该方法在原YOLOv5主干网络添加屏蔽卷积注意力模型,该模块从一个屏蔽卷积层开始,感受野的中心区域被遮掩,通过预测屏蔽信息并利用与屏蔽信息相关的误差作为异常得分。在检测网络中嵌入Swin-CA模块。通过对相邻层特征的学习,使得模型能够更好地掌握全局信息,从而减小了背景信息对检测结果的影响,通过提取不同背景中人体异常行为尺度特征,降低了整个模型计算的复杂度,提高了模型对人体异常行为目标定位的精度。在UCSD-ped1、KTH和Shanghai Tech数据集上的实验结果表明,提出方法的检测精度分别达到了98.2%、96.4%和95.8%。

关键词:人体异常行为;YOLOv5;屏蔽卷积;注意力机制;Swin-CA模块
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.2208-0061

Optimizing Human Abnormal Behavior Detection Method of YOLO Network

ZHANG Hongmin[1,2], ZHAUNG Xu[1], ZHENG Jingtian[1], FANG Xiaobing[1]
1.School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China
2.Liangjiang International College, Chongqing University of Technology, Chongqing 401135, China

Abstract: Because of the large interference of environmental background information in public surveillance videos and the different scale of abnormal human behavior goals, at present, it is difficult to improve the precision of human abnormal behavior detection. For the above issues, this paper designs the abnormal behavior detection method by improving the YOLOv5 module. In this method, a shielded convolutional attention model is added to the original YOLOv5 backbone network. The module starts from a shielded convolutional layer, and the central region of the receptive field is covered. The shielding information is predicted and the errors related to the shielding information are used as abnormal scores. At the same time, Swin-CA module is embedded in the detection network. Through the study of characteristics of adjacent layers, enables the module to get stronger grasp the overall situation information, thus reducing the affect of backdrop message on the detection results, by extracting the scale characteristics of human behavior abnormalities in different backgrounds, it decreases the order of complex of the whole model calculation and improves the precision of the module to locate the target of abnormal human behavior. Experimental results on the UCSD-PED1, KTH and Shanghai Tech datasets show that the precision of the proposed method reaches 98.2%, 96.4% and 95.8%, respectively.

Key words: abnormal human behavior; YOLOv5; mask convolution; attentional mechanism; Swin-CA module

目前,视频监控系统普遍应用于公共场所,在社会治安管理、打击犯罪、城市管理、服务人民生活等领域发挥着重要作用[1-2]。然而现实生活中人体异常行为类型复杂、数量众多,且不同的情况下非正常行为的界定标准也不一样[3]。大多数情况下异常行为样本仅在测试的时候可用,因此原YOLO网络不适合直接应用于人体异常行为检测。

部分研究人员把目光投向了其他技术上,如基于重建的方法[4-5]、字典学习方法[6-7]等。重建方法的一个显著特征就是子类别依赖于预测掩蔽信息,利用相对于掩蔽

基金项目:重庆市自然科学基金面上项目(cstc2021 jcyj-msxmX0525)。
作者简介:张红民(1970—),通信作者,男,博士,教授,主要研究方向为图像处理与模式识别,E-mail:hmzhang@cqut.edu.cn;庄旭(1999—),男,硕士研究生,主要研究方向为信号与信息处理;郑敬添(2000—),男,硕士研究生,主要研究方向为信号与信息处理;房晓冰(1995—),男,硕士研究生,主要研究方向为信号与信息处理。
收稿日期:2022-08-04 **修回日期:**2022-11-10 **文章编号:**1002-8331(2023)07-0242-08

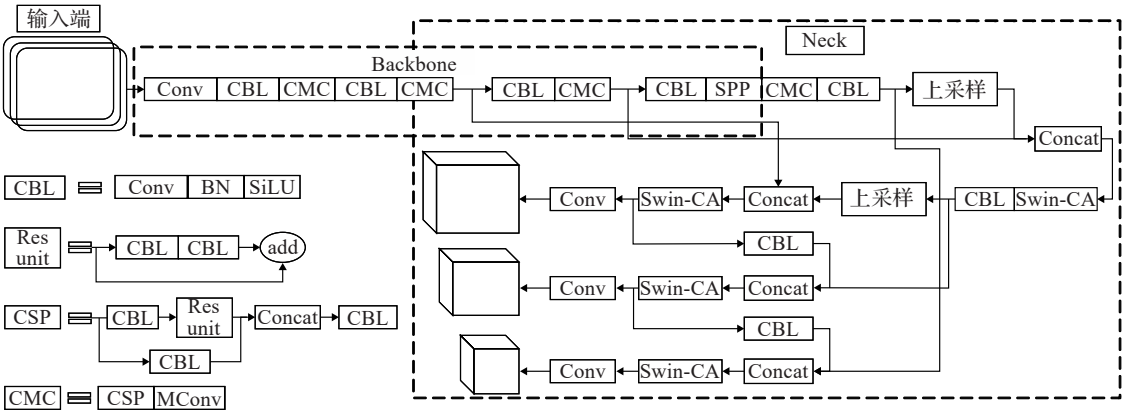


图1　改进后整体模型结构图

Fig.1　Improved overall model structure diagram

信息的重建误差作为异常得分。同时,目前公共场合监控视频数据中人体目标的尺度不同为当前人体异常行为检测的准确率以及检测速度带来了难题。为了解决目标的尺度不同这个问题,Lin等[8]于2017年首次公开地给出了FPN特征金字塔网络来处理目标分类中的多尺度问题,该技术通过利用最简单的网络连接改变,从而使得对于小对象的检测和分析性能较大提高,但是,通过特征提取和将不同层次的特征加以融合的做法并不能将更多的特征集合在一起。

为了能够更好地提取多尺度特征并将其融合,研究人员对特征金字塔网络进行改进,并于2018年提出了PAN网络结构[9],PAN网络虽然有效地解决了特征融合的问题,但是对人体目标的检测及计算速率十分缓慢。

文献[10]提出了YOLOv4。YOLOv4采用了FPN和PAN相结合的网络结构,使图像的视觉范围得到了最大程度的提高,同时也能迅速地分割出更加重要的特征信息。YOLOv5[11]沿用了v4版本的网络结构,继续采用FPN+PAN结构,但是经过实验发现FPN+PAN结构并不能很好地提取人体异常行为目标的特征,同时在上采样的过程中还会产生重叠效应。为此,本文对YOLOv5的网络结构进行改进,提出了一种人体异常检测方法(MCS-YOLO)。

1　优化YOLO网络的人体异常行为检测方法

1.1　MCS-YOLO方法改进思路

YOLOv5模型在目标检测领域中取得了很好的效果,但对于具有复杂背景的人体异常行为的图片,YOLOv5很难得到比较精确的特征,容易产生误检或者漏检。此外,YOLOv5模型无法检测各种尺度的行为对象。

针对上述问题,本文对YOLOv5网络结构进行分析,对YOLOv5模型优化得到新的人体异常行为检测模型:MCS-YOLO。本文利用重建方法的思想,通过在主干网络中添加屏蔽卷积注意力模块MC,该模块由一个屏蔽卷积层[12]和注意力机制组成,屏蔽卷积层基于一个自定义的感受野,在那里图片的中心区域被屏蔽。产生的卷积激活图随后通过注意模块传递。注意模块用来确保网络不会简单地学习基于线性插值上下文信息的屏蔽区域,对网络的性能带来极大的改进。在YOLOv5网络的FPN+PAN结构中嵌入Swin-CA模块。Swin-CA模块由Swin transformer模块与坐标注意力结合而成,利用Swin transformer模块来引导特征聚合以改进多尺度特征学习的方式,借助坐标注意力机制的特点来获取特征图片的精确的位置信息。Swin transformer可以从各种尺度中提取出异常的人类行为[13],能够提高模型的目标识别能力。坐标注意力机制[14]能够精准地定位检测目标的位置。因此在本文中,将Swin transformer模块与坐标注意力机制结合,让模型能够得到更加丰富的图片的特征信息。该方法对原YOLOv5网络的主要改进有以下几点:首先,在YOLOv5主干网络中添加屏蔽卷积注意力模块MC,提高网络的特征学习能力,优化网络的检测性能;然后,在YOLOv5中的检测网络中嵌入Swin-CA模块,提取人体目标多尺度特征,利用坐标注意力机制,可以准确地保存人类行为对象的位置信息。优化后的MCS-YOLO网络如图1所示。

1.2　屏蔽卷积注意力模块

本文介绍了一种屏蔽卷积,如图2所示。利用该卷积进行上下文信息学习从而预测隐藏信息。同时将该卷积应用到注意力机制中,即屏蔽卷积注意力模块MC。

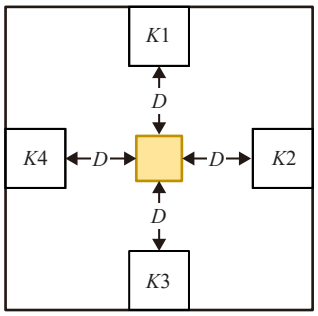


图2　屏蔽卷积

Fig.2　Masked convolutional

屏蔽卷积注意力模块的作用是利用上下文信息学习重建隐藏信息。为了实现这一结果,本文将模块设计为一个带有扩展的屏蔽卷积层,然后是通道注意模块。在屏蔽卷积注意力模块中设计了关于预测信息的损失函数,其目的是最小化被屏蔽输入和预测输出之间的重建误差。

1.2.1　屏蔽卷积

屏蔽卷积的感受野如图2所示。该卷积的可学习参数位于感受野的角落 $\boldsymbol{K}_i \in \mathrm{R}^{k'\times k'\times c}$，$\forall i \in \{1,2,3,4\}$ 表示感受野4个角落的卷积核,其中 $k' \in N^+$ 是定义子内核大小的超参数，c 是通道数。每个核 K_i 位于距离感受野中心的掩蔽区域 $D \in N^+$ 的距离,用 $\boldsymbol{M} \in \mathrm{R}^{1\times 1\times c}$ 表示感受野中心的掩蔽区域。因此,感受野的空间大小 k 计算如下:

$$k = 2k' + 2D + 1 \tag{1}$$

$\boldsymbol{X} \in \mathrm{R}^{h\times w\times c}$ 是屏蔽卷积层的输入张量，h 和 w 分别是高度和宽度。在输入 X 的某个位置使用自定义核执行的卷积运算只考虑子核 K_i 所在位置的输入值,而忽略其他信息。每个 K_i 和相应输入之间的卷积运算结果被求和为一个数字。结果值表示与位于 M 相同位置的预测。一个屏蔽卷积产生单个激活图,为了预测 M 中每个信道的值,引入 c 个屏蔽卷积,每个卷积预测来自不同信道的屏蔽信息。由于该模块的目标是学习和预测输入的每个空间位置的重建,本文在输入周围添加了 $k'+D$ 像素的零填充,并将步幅设置为1,这样输入中的每个像素都被用作屏蔽信息。因此,输出张量 Z 的空间尺度与输入张量 X 的空间尺度一致。最后,输出张量通过ReLU激活。

1.2.2　通道注意模块

接下来,屏蔽卷积的输出由通道注意模块处理,该模块计算每个通道的注意分数。输出张量 Z 中的每个激活图都是在存在掩蔽信息的情况下由单独的屏蔽卷积预测的,由此可以推断屏蔽卷积最终会生成包含有不成比例的跨通道值的激活图,即会得到通道之间的关系。利用文献[15]所提供的通道注意机制对通道的特性响应可以进行自适应校正,而利用这个机制,系统在特征提取时就能够提前使用全局图像信息,在必要时还能选择性强调或抑制重建信息。而使用注意力的另一原因是用于对屏蔽卷积注意力模块的输入和输出关系的非线性处理。

通道注意模块通过在每个通道上执行全局池化将张量 Z 减少为向量 $z \in \mathrm{R}^c$。然后,计算比例因子 $s \in \mathrm{R}^c$,计算如下:

$$s = \sigma(\boldsymbol{W}_2 \cdot \delta(\boldsymbol{W}_1 \cdot z)) \tag{2}$$

其中 σ 是Sigmoid激活，δ 是ReLU激活，$\boldsymbol{W}_1 \in \mathrm{R}^{(c/r)\times c}$ 和 $\boldsymbol{W}_2 \in \mathrm{R}^{c\times(c/r)}$ 分别表示两个连续的完全连接层(FC)的权重矩阵。第一个FC层由 c/r 个神经元组成,以 r 的还原率压缩信息。然后,在空间维度中复制向量 s,生成与 Z 大小相同的张量 S。最后一步是 S 和 Z 之间的元素相乘,生成最终包含重新校准特征图的信息的张量 $\hat{\boldsymbol{X}} \in \mathrm{R}^{h\times w\times c}$。

1.3　添加Swin-CA模块的检测网络

目前人体异常行为检测面临检测图像的尺寸变化大和系统运算复杂度高的挑战。为使检测网络具备较高检测速率的同时,进一步提高检测准确度,更好地使用在检测网络中的特征信息,在YOLOv5的检测网络中引进了由Swin transformer模型与坐标注意力机制相结合的新模型,即Swin-CA。

1.3.1　坐标注意力机制

本文在对注意力机制的研究中发现一般的注意力模型会忽略对人体异常行为至关重要的位置信息。为此本文采用了一种基于坐标的注意力机制,它在一定的空间方位上捕捉目标的位置知觉依赖关系,然后在其他的具体目标的空间方位上,保留了更完整、更精确的目标的空间定位信息,从而产生了具有空间方位感知相关特征的特征图,通过补充和有效地利用要输入此特征的特征图,以增强感兴趣的目标特征的表示。通过引入坐标注意力机制,在YOLOv5模型的检测过程中能够有效的捕获通道之间的关系,保留目标的确切位置,使网络能够更准确地识别目标并提高检测精度,同时在计算方面避免了大量的计算开销。

1.3.2　Swin-CA模块

为了降低注意力机制的计算复杂度,同时提取多尺度特征,本文将Swin transformer模块与坐标注意力机制相结合,组成Swin-CA模块,并将其嵌入到YOLOv5的检测网络中。Swin transformer采用了分组运算的思想,通过采用CNN结构中常见的分层构造方法来实现各特征矢量的融合,使模块可以掌握全局信息,而在特征图中引入坐标注意力机制,则可以更好地利用特征图中的异常行为对象的位置信息。Swin-CA模块结构如图3所示。

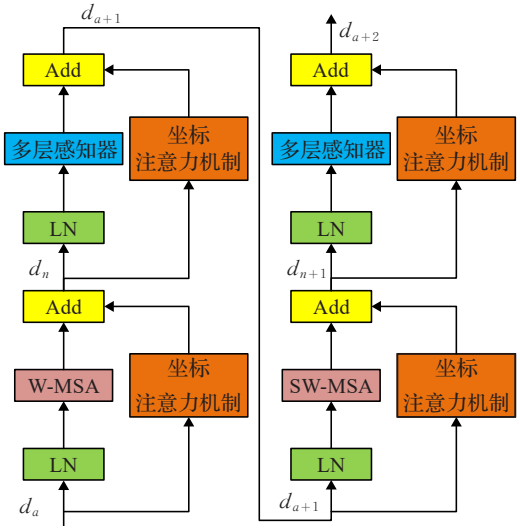


图3　Swin-CA模块

Fig.3　Swin-CA module

Swin transformer模块主要由窗口多头自我注意层模块和移位窗口多头自我注意层模块分别组成。将屏蔽卷积注意力模块计算出的张量 $\hat{X}\in \mathrm{R}^{h\times w\times c}$(图中表示为 d_a)输入到窗口多头自注意层模块中,通过W-MSA模块进行特征学习并进行残差运算,同时将输入 d_a 送入坐标注意力机制并将两者的值相加得到输出特征 d_n。输出特征 d_n 再分别经过坐标注意力机制和一个LN层和MLP层,最后进行残差运算,得到输出 d_{a+1}。公式如式(3),(4)所示。

$$d_n = W-MSA(LN(d_a))+CA(d_a) \tag{3}$$

$$d_{a+1} = MLP(LN(d_n))+CA(d_n) \tag{4}$$

移位窗口多头自我注意层的结构与窗口多头自我注意层的结构类似,唯一不同的是该层是利用SW-MSA模块来计算图片的特征部分。

将Swin-CA模块嵌入到YOLOv5的检测网络中,可以让模型更好地掌握全局信息,借鉴CNN网络分层构造方法,将所抽取的特征进行多尺度的划分,从而极大地减少了计算过程的复杂性。

1.4 损失函数

为了充分利用屏蔽卷积注意力模块的特性,本文在对人体异常行为检测之外添加了一个监督任务,该任务包括屏蔽卷积感受野在内的每个屏蔽卷积的位置重建屏蔽区域。为此,屏蔽卷积注意力模块为每一个屏蔽区域提供相应的重建作为输出 $\hat{X}$。令 F 表示MC模块,将监督任务的重建损失定义为输入和输出之间的均方误差,如下所示:

$$L_{\mathrm{MC}} = (\hat{X}-X)^2 \tag{5}$$

将该损失函数的值简单地添加到YOLOv5网络的原损失数值中,从而产生一个新的损失函数,该函数包含两个项:

$$L_{\mathrm{total}} = L_Y + \lambda L_{\mathrm{MC}} \tag{6}$$

式中 $\lambda\in \mathrm{R}^+$ 是一个超参数,它用来控制 F 对于整体损失函数的重要性,而 L_Y 是原YOLOv5网络的损失函数。

2 实验结果与数据分析

2.1 实验数据集与参数设置

实验在UCSD-ped1[16]、KTH[17]和Shanghai Tech[18]3个公共的人体异常行为数据集上进行。UCSD-ped1数据集包含70个人类行为的视频,视频来自室外场景,使用静态摄像头以每秒10帧的速度录制。在这些视频场景中的主要移动对象是行人,即正常行为。因此,所有其他物体(如汽车、滑板、轮椅或自行车)都被视为异常行为。Shanghai Tech数据集包含了13台高清摄像机拍摄的330个正常的动作和107个不正常的动作,该数据集中包含11种不同的人体行为,比如骑自行车、溜冰、打架、抢劫、摔倒等,每一个视频都有856×480的清晰度。KTH数据集与之前两个数据集有所不同,数据集中只包含了6种动作,数据集中视频的场景分为室内和室外。

本文的实验平台是pytorch框架。网络输入图片尺寸按照数据集的不同分别编辑为相应图片大小,初始训练的学习率 lr 设置为0.005,图片批数量设置为8,每一个数据集都训练150个epoch。对模型训练时,利用迁移学习来加快模型的训练速度,将原YOLOv5网络的权重文件作为MCS-YOLO网络的初始训练权重,极大地较少了模型训练时间并得到了良好的检测结果。

为了验证本文所提出的方法的有效性,选择准确率(Auc)、平均精度(mAp)、损失函数(Loss)以及模型运行测试集所耗费的时间等指标。

其中准确率(Auc)是为了评价本文方法在分类效果上的好坏。在使用YOLO方法进行人体异常行为检测时,考虑到与正常行为差异较大的都应被定义为异常行为,因此本文添加准确率(Auc)作为MCS-YOLO的评价指标之一。对MCS-YOLO训练之前需要对异常行为进行标记,本文将含有标记的帧图片定义为负样本,没有标记的图片定义为正样本。在测试时采用数据集中的所有帧图片进行测试并计算Auc值。具体计算如下:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

式中每部分代表的含义如表1所示。

表1 Auc各参数含义

Table 1 Meaning of each parameter in Auc

	正样本(P)	负样本(N)
预测正确(T)	TP	TN
预测错误(F)	FP	FN

平均精度(mAp)代表在数据集每个类别的平均精度的平均数。mAp越高则表示模型对于各类别的平均检测效果越好。损失函数(Loss)是指用于衡量模型的预测值与实际数值之间的不同程度的运算函数。

2.2 MCS-YOLO方法通用性实验及结果

本文中对MCS-YOLO方法在通用条件下的目标检测特性进行了研究。在PASCAL VOC数据集上对MCS-YOLO方法进行了检验。该数据集总共分为20种类别,总计约18 000张图片。从实验开始就划分了这个数据集,并根据3:1的比例分成了训练集和验证集,选择mAp(mean average precision)作为评价的标准。与4种目标检测的方法比较,比较结果如表2所示[19-21]。

表2 PASCAL VOC数据集下的实验结果

Table 2 Experimental results under PASCAL VOC dataset

方法	mAp/%	时间/s
YOLOv3[19]	78.2	330
YOLOv3-MSEE[20]	81.2	340
YOLOv4[21]	82.7	303
YOLOv5	83.4	310
MCS-YOLO	85.7	320

从表2中的结果可以看得出,MCS-YOLO方法相较于YOLOv5提高了2.3个百分点,在运行验证集方面比YOLOv5增加了10 s。由此可以看出,MCS-YOLO方法的检测性能要优于YOLOv5,在检测速度方面也与原YOLOv5模型相差不多,速度能够得到保证。实验的结果表明MCS-YOLO在特征提取方面有了显著提升,对于图片中的特征信息掌握得更加全面,由于在检测网络添加了Swin-CA模块,使得MCS-YOLO方法在图片的多尺度检测性能方面得到了较大的提升。因此,可以推断MCS-YOLO方法适合用于对人体异常行为的检测。

2.3　与目前研究方法比较

表3中展示了MCS-YOLO与其他方法的比较,主要是比较不同方法之间的Auc值[5,22-25]。从表中可以看出,本文方法相较于RGB-STCNN和Two-Stream I3D这两种方法,在KTH数据集上的Auc提升明显,分别提高了6.9个百分点和23.4个百分点。在UCSD-ped1数据集上,相较于其他方法也有提升,说明MCS-YOLO能够更好地提取图片中的特征信息,对于图片中的重建信息也进行了很好地利用。在Shanghai tech数据集上,MCS-YOLO测试得到的Auc值略低于Lu等人提出的MAML方法,Shanghai tech数据集在校园拍摄,来往人员复杂且有重叠等问题导致在提取异常行为特征时会受到干扰,从而降低了模型对异常行为的分辨能力。

表3　不同方法在3种数据集上的Auc对比

Table 3　Auc comparison of different methods of three datasets

单位:%

方法	UCSD	Shanghai tech	KTH
RGB-STCNN[22]	—	—	89.8
Two-Stream I3D[22]	—	—	73.3
ConvAE[5]	85.0	60.9	—
AbnormalGAN[23]	93.5	—	—
Dong et al[24]	95.6	73.7	—
Lu et al[26]	96.2	77.9	—
MCS-YOLO	96.9	75.5	96.7

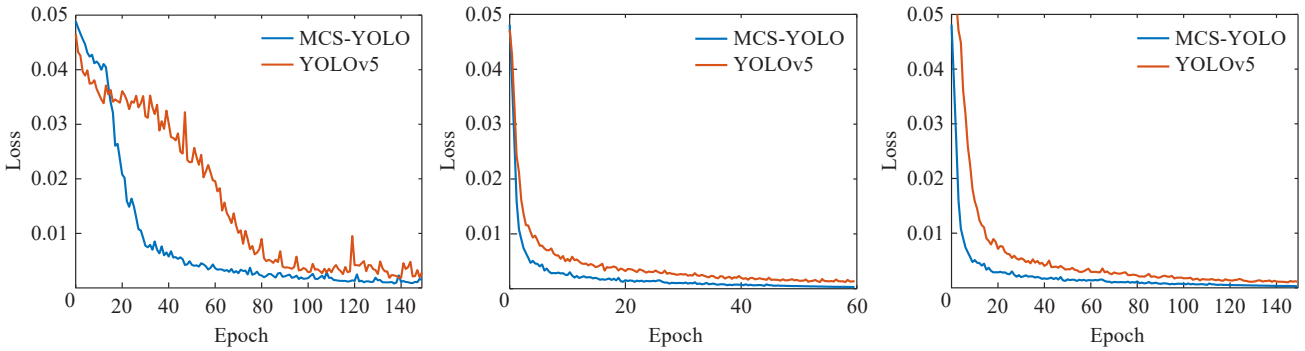
2.4　不同损失函数的对比曲线图

如图4是在3个数据集上分别采用原YOLOv5损失函数与添加了重建损失的新损失函数,在同一模型MCS-YOLO下的损失值随训练轮次的变化的对比结果图。红色表示采用原损失值的变化情况,蓝色表示添加重建损失的新损失函数的变化情况。图片中的横坐标表示训练轮次epoch,纵坐标代表损失值。从图中可以看出,本文采用的添加重建损失的损失函数的初始损失值和原YOLOv5的初始损失值相差不大;在UCSD-ped1数据集上当训练次数达到30次后MCS-YOLO开始收敛,而原YOLOv5模型在epoch达到90次左右的时候才开始收敛。在KTH和Shanghai Tech数据集上,两种方法的曲线大致相同。

如表4展示了采用YOLOv5损失函数的MCS-YOLO和添加重建损失的MCS-YOLO在3个数据集上的检测结果对比,从表中可以看出在YOLOv5模型中加入重建损失对模型检测的mAp值提升较小,但是添加重建损失的MCS-YOLO经过较短时间的训练就能迅速收敛比采用YOLOv5损失函数的MCS-YOLO收敛速度更快,且最终都能收敛在较低的损失值。

2.5　实验结果对比分析

上述选择的3个人体异常行为的数据集本身是视频,本文首先对视频进行处理,将其分为帧图片再对不同类别的人体异常行为进行标注。由于数据集的制作中有部分视频拍摄较为模糊且部分特征被背景遮挡,本



(a)UCSD-ped1数据集　(b)KTH数据集　(c)Shanghai Tech数据集

图4　3种网络架构训练的Loss对比

Fig.4　Loss plots trained on three network architectures

表4　两种损失函数对比

Table 4　Comparison of two loss functions

损失函数	mAp/%			速度/(frame/s)		
	UCSD-ped1	KTH	Shanghai tech	UCSD-ped1	KTH	Shanghai tech
MCS-YOLO	98.6	95.2	96.8	40.930	42.152	43.450
YOLOv5	96.3	94.5	95.3	42.450	44.182	45.150

表5　3种模型的性能对比

Table 5　Performance comparison of three models

方法	mAp/%			速度/(frame/s)		
	UCSD-ped1	KTH	Shanghai tech	UCSD-ped1	KTH	Shanghai tech
MCS-YOLO	98.6	95.2	96.8	40.930	42.152	43.450
YOLOv5	90.8	91.5	92.3	39.450	43.182	42.150
YOLOv4	89.5	89.6	90.7	40.240	42.117	43.750

(a)UCSD-ped1数据集　(b)KTH数据集　(c)Shanghai tech数据集

图5　3种网络架构的检测结果对比

Fig.5　Comparison of detection results of three network architectures

文在制作实验数据集时通过筛选选取高质量的图片。将图片格式统一转化为jpg,并通过make sense工具对数据集中人体异常行为特征进行标注,生成了TXT标签文档。

本文分别在YOLOv4、原YOLOv5网络和MCS-YOLO方法上对制作的数据集进行训练,然后对检测结果中的mAp和检测所耗费的时间这些指标进行统计,测试的结果如表5所示。测试的结果表明:优化后的MCS-YOLO在UCSD-ped1上提升最为明显,mAp值达到了98.6%,相对于YOLOv5和YOLOv4分别提高了7.8个百分点和9.1个百分点,运行验证集所花费时间为每秒40.93 frame,对比YOLOv4和YOLOv5的运行速度有所降低。而在另外两个数据集上MCS-YOLO方法相较于YOLOv5网络的mAp值分别提高了3.7个百分点、4.5个百分点左右,相较于YOLOv4提升了5.6个百分点、6.1个百分点。对于运行验证机耗费的时间来说MCS-YOLO方法的提升不大。总体来说,MCS-YOLO方法在检测精度上得到了很大的提升,并且在验证集上的检测速度也有一定的保证。如图5是3种模型的检测效果的对比示例,其中第一行到第三行,分别表示了MCS-YOLO、YOLOv5以及YOLOv4模型的检测结果,第一列到第三列则分别表示UCSD-ped1数据集、KTH数据集、Shanghai tech数据集中的检测结果。

从表5的指标结果及图5结果图中可以看出,在UCSD-ped1数据集和Shanghai tech数据集上,每种方法都能够检测出滑冰这一异常行为,其中MCS-YOLO模型检测的置信度最高,其次是YOLOv5模型,最低的是YOLOv4模型。而在KTH数据集中,MCS-YOLO与YOLOv5模型的检测结果并没有太大区别,而YOLOv4模型对跑步这一异常行为的检测置信度较低。从整体的检测结果分析,MCS-YOLO相较于YOLOv5模型的人体异常行为的检测性能更强,能够更准确地检测出更多的异常行为目标。

为了进一步验证MCS-YOLO网络检测人体异常行为的有效性,本文对3个数据集中的人体异常行为进行了分类并对网络进行训练,训练结果如表6~8所示。从表中的mAp值的结果来看,MCS-YOLO相较于YOLOv5与YOLOv4网络,在数据集上的分类性能更优,能够更

表6　UCSD-ped1数据集下的检测结果

Tab 6　Detection results under UCSD-ped1 dataset

单位:%

Category	MCS-YOLO	YOLOv5	YOLOv4
cart	99.5	92.3	92.0
wheelchair	99.5	94.6	94.4
skater	95.2	87.8	87.3
through the lawn	99.5	89.5	90.2
bike	99.5	89.8	88.8

加精确地检测出不同类型的人体异常行为,这也表明MCS-YOLO网络可以适用于不同场景下的人体异常行为的检测。

表7　KTH数据集下的检测结果

Table 7　Detection results under KTH dataset

单位:%

Category	MCS-YOLO	YOLOv5	YOLOv4
jogging	92.1	89.4	89.2
running	98.3	93.6	93.3
boxing	96.7	92.8	92.3
hand waving	92.8	89.5	88.9
hand claping	96.1	92.2	91.9

表8　Shanghai tech数据集下的检测结果

Table 8　Detection results under Shanghai tech dataset

单位:%

Category	MCS-YOLO	YOLOv5	YOLOv4
bike	97.5	97.1	96.5
fight	96.1	95.5	94.3
car	96.9	95.8	96.0
skate	95.3	94.5	94.1
through a bag	98.2	97.8	97.4
motor	97.2	96.9	97.3
jump	98.1	97.7	97.8
run	97.6	97.3	96.9
tumble	94.8	94.6	95.8
robbery	96.3	92.3	91.8

2.6　消融实验

本文为进一步验证MCS-YOLO方法对人体异常检测的有效性,通过进行消融实验分析各个优化点对与YOLOv5的改进效果,选择UCSD-ped1数据集来进行该实验。实验结果如表9所示,分别添加屏蔽卷积注意力模块、Swin-CA模块以及重建损失,每个模块都不同程度地提升了模型的整体性能。

表9　消融实验

Table 9　Ablation experiment

屏蔽卷积注意力模块	Swin-CA模块	重建损失	mAp/%
×	×	×	90.8
√	×	×	96.1
×	√	×	93.5
√	×	√	96.3
√	√	√	98.6

在原YOLOv5的主干网络中引入屏蔽卷积注意力模块,提升模型掌握全局信息的能力,网络的特征提取能力提升,人体异常目标的检测准确率有了较大的提升,有效地解决了原YOLOv5在进行人体异常行为检测时准确率不高的问题。引入屏蔽卷积注意力模块的YOLOv5网络在检测数据集中每一类异常行为的准确率都有较大的提升,相比原YOLOv5的mAp值提高了5.3个百分点。

在检测网络中嵌入Swin-CA模块,mAp值提高了2.7个百分点。Swin-CA模块强化了模型对于不同大小的人体异常行为的检测能力,通过添加坐标注意力机制提升了网络对异常行为的定位能力,改善了网络对于多尺度异常行为的特征提取效果,mAp值提高了2.7个百分点。

在YOLOv5模型中加入重建损失对模型的整体性能提升并没有太大的作用,但从上述对于损失函数的分析可以看出,添加重建损失可以加快模型的收敛的速度。

从消融实验的对比结果来看,MCS-YOLO相较于YOLOv5,检测速率没有太多降低,但是检测的准确率却有了一个大幅度的提升,进一步说明了MCS-YOLO方法的有效性。

3　结语

本文提出的MCS-YOLO方法的主要创新点:(1)在YOLOv5网络结构中加入了屏蔽卷积注意力模块,提高了模型的特征提取能力以及检测网络的准确率;(2)在原有检测网络的基础上引入Swin transformer模块以及坐标注意力机制提高了人体异常行为的特征表现能力,提高了检测网络的准确性。该方法对检测人体异常行为具有积极意义。

不过,MCS-YOLO方法还存在着不足:(1)由于不同数据集所收集的情景不同,对人体异常行为的定性在不同的情景中也是有所不同的,使得该方法通用性较为欠缺;(2)对模型的训练需要提前检测图像进行标注,前期工作量较大;(3)该方法对于图像中人动作的连续性并不敏感,使得检测过程中对于人体异常情况的判断出现相应的延迟或误检、漏检。

参考文献:

[1] LENTZAS A, VRAKAS D.Non-intrusive human activity recognition and abnormal behavior detection on elderly people: a review[J].Artificial Intelligence Review, 2020, 53(3):1975-2021.

[2] ZHANG X P, JI J H, WANG L, et al.Review of video based human abnormal behavior recognition and detection[J].Control and Decision, 2021(1):1-14.

[3] FAN Z, YIN J, SONG Y, et al.Real-time and accurate abnormal behavior detection in videos[J].Machine Vision and Applications, 2020, 31(7):1-13.

[4] DONG G, LIU L Q, LE V, et al.Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection[C]//Proceedings of International Conference on Computer Vision, 2019: 1705-1714.

[5] HASAN M, CHOI J, NEUMANN J, et al.Learning temporal regularity in video sequences[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Rec-

ognition(CVPR),2016.

[6] CARRERA D,MANGANINI F,BORACCHI G,et al. Defect detection in nanostructures[J].IEEE Transactions on Industrial Informatics,2017,99:1.

[7] CHENG K W,CHEN Y T,FANG W H.Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2015.

[8] LIN T Y,DOLLAR P,GIRSHICK R,et al.Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017.

[9] HAN C L,PENG F X,JIE A,et al.Pyramid attention network for semantic segmentation[J].arXiv:1805.10180,2018.

[10] GUO S,ZHONG P,SUN Y,et al.Fast detection algorithm for surface defects of metal parts based on YOLOv4-mobilenet network[C]//Proceedings of Society of Photo-Optical Instrumentation Engineers(SPIE),2021.

[11] LIU R,WANG H,ZHANG S,et al.Object detection algorithm based on improved YOLOv5 for basketball robot[C]//Proceedings of Chinese Intelligent Systems Conference,2022.

[12] RISTEA N C,MADAN N,IONESCU R T,et al.Self-supervised predictive convolutional attentive block for anomaly detection[J].arXiv:2111.09099,2021.

[13] LIU Z,LIN Y,CAO Y,et al.Swin transformer:hierarchical vision transformer using shifted windows[C]//Proceedings of International Conference on Computer Vision (ICCV),2021:9992-10002.

[14] HOU Q,ZHOU D,FENG J.Coordinate attention for efficient mobile network design[C]//Proceedings of International Conference on Computer Vision and Pattern Recognition(CVPR),2021:13708-13717.

[15] HU J,SHEN L,ALBANIE S,et al.Squeeze-and-excitation networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(8):2011-2023.

[16] MAHADEVAN V,LI W X,BHALODIA V,et al.Anomaly detection in crowded scenes[C]//Proceedings of Computer Vision & Pattern Recognition,2021.

[17] SCHULDT C,LAPTEV I,CAPUTO B.Recognizing human actions:a local SVM approach[C]//Proceedings of International Conference on Pattern Recognition,2004:32-36.

[18] LUO W,WEN L,GAO S.A revisit of sparse coding based anomaly detection in stacked RNN framework[C]//Proceedings of 2017 IEEE International Conference on Computer Vision(ICCV),2017:341-349.

[19] 张富凯,杨峰,李策.基于改进YOLOv3的快速车辆检测方法[J].计算机工程与应用,2019,55(2):12-20.
ZHANG F K,YANG F,LI C.Fast vehicle detection method based on improved YOLOv3[J].Computer Engineering and Applications,2019,55(2):12-20.

[20] 张红民,李萍萍,房晓冰,等.改进YOLOv3网络模型的人体异常行为检测方法[J].计算机科学,2022(4):233-238.
ZHANG H M,LI P P,FANG X B,et al.Human abnormal behavior detection method based on improved YOLOv3 network model[J].Computer Science,2022(4):233-238.

[21] 徐印赟,江明,李云飞,等.基于改进YOLO及NMS的水果目标检测[J].电子测量与仪器学报,2022,36(4):114-123.
XU Y Y,JIANG M,LI Y F,et al.Fruit target detection based on improved YOLO and NMS[J].Journal of Electronic Measurement and Instrumentation, 2022, 36(4):114-123.

[22] MARTIN P E,BENOIS-PINEAU J,R PÉTERI,et al.3D convolutional networks for action recognition:application to sport gesture recognition[J].arXiv:2204.08460,2022.

[23] RAVANBAKHSH M,NABI M,SANGINETO E,et al. Abnormal event detection in videos using generative adversarial nets[C]//Proceedings of 2017 IEEE International Conference on Image Processing(ICIP),2017:1577-1581.

[24] DONG F,ZHANG Y,NIE X S.Dual discriminator generative adversarial network for video anomaly detection[J]. IEEE Access,2020,8:88170-88176.

[25] LU Y W,YU F,REDDY M K K,et al.Few-shot scene-adaptive anomaly detection[C]//Proceedings of European Conference on Computer Vision,2020:125-141.