◆特邀栏目◆

基于双尺度串行网络的视频异常行为检测*

吴德刚[1],赵利平[2**],陈乾辉[1],张宇波[3]

(1.商丘工学院机械工程学院,河南商丘 476000;2.商丘工学院信息与电子工程学院,河南商丘 476000;3 郑州大学电气与信息工程学院,河南郑州 450000)

摘要:针对传统视频异常行为检测模型存在的性能不佳与时间开销较大的问题,从空间和时序维度构造双尺度串行网络的视频异常行为检测模型(Dual-Scale Serial Network,DSS-Net)。首先,利用深度可分离卷积对Vgg-16网络进行改进,并利用改进的特征提取器从空间维度提取特征,从而可以通过减少计算参数量来降低模型的时间开销。接着,在此基础上引入注意力机制,从而强化目标特征的表达能力。最后,利用长短期记忆(Long Short-Term Memory,LSTM)网络从时序维度提取运动视频每一帧之间的上下文时序关系。在当前主流的UCSD Ped1和Ped2数据集以及更具挑战性的UCF数据集上进行测试,结果表明,在3个数据集上DSS-Net的ROC(Receiver Operating Characteristic)线下面积(Area Under Curve,AUC)值分别达到95.30%、96.80%、80.60%,等错误率(Equal Error Rate,EER)分别达到10.60%、12.60%、18.50%,同时具有更强的实时性。相比经典的One-class Neural Network (ONN)和Aggregation of Ensembles (AOE)模型,DSS-Net在Ped1和Ped2数据集上的AUC值分别提升了0.42%和0.94%。此外,DSS-Net也在UMN、ShanghaiTech和CUHK Avenue等数据集上进行了泛化能力和鲁棒性的测试,结果与当前主流模型相比具有一定的竞争力。

关键词:视频异常行为检测;空间维度;时序维度;深度可分离卷积;注意力机制

中图分类号:TP393 **文献标识码**:A **文章编号**:1005-9164(2023)03-0575-12

DOI:10.13656/j.cnki.gxkx.20230710.017

随着安防系统的不断完善,由公共摄像头组成的庞大监控网络已经在我国形成。基于监控视频进行异常行为检测已成为计算机视觉中的一个重要研究领域[1],并在学校、医院、交通、大型商场等实时监控系统中得到广泛的应用。

传统的人工异常行为检测包括多个步骤[2],如视

收稿日期:2022-10-31 修回日期:2023-01-09

*河南省高等学校青年骨干教师培养计划项目(2018GGJS190)和商丘工学院2022科研项目(2022KYXM02)资助。

【第一作者简介】

吴德刚(1980-),男,副教授,主要从事异常事件检测、机器视觉研究。

【**通信作者】

赵利平(1981-),女,副教授,主要从事智能控制、图像检测研究,E-mail:zhaolp3360@sina.com。

【引用本文】

吴德刚,赵利平,陈乾辉,等.基于双尺度串行网络的视频异常行为检测[J].广西科学,2023,30(3):575-586.

WU D G,ZHAO L P,CHEN Q H,et al. Video Abnormal Behavior Detection with Dual-Scale Serial Network [J]. Guangxi Sciences,2023,30(3):575-586.

频采集和异常行为识别与分析。由于检测阶段需要大量的人力和物力,且当监控人员疏忽时极易造成检测遗漏等问题。因此,越来越多的研究人员开始关注异常行为自动检测,并利用不同的计算机视觉方法对异常行为进行建模、识别、分析以及快速定位。

传统的异常行为检测方法主要分为动态贝叶斯网络、概率主题模型、集群模型和稀疏表示法[3,4]。其中,隐马尔可夫模型及其变体是所有动态贝叶斯网络中最常用的异常行为检测方法。动态贝叶斯网络一般将人类行为表示为一组状态向量,并将概率作为行为模式和测试序列的相似性。此外,期望最大化算法、最大概率估计和最大后验估计分类器等方法也被用来检测异常行为。为了降低动态贝叶斯网络的巨大计算成本,许多工作提出使用概率主题模型进行异常行为检测,这包括隐含概率语义模型和隐含分析模型两大类[5]。其中,基于隐含概率语义模型的异常行为检测方法通过推断隐藏变量的潜在分布来查找异常特征分布,并将出现概率较低的特征视为异常特征。此外,聚类模型,包括 K-means、Markov Cluster (MCL)等方法也常被用于异常行为的检测[6],正常行为通常被聚集成几个聚类,这些聚类分别代表不同的正常行为模式,而远离这些聚类的行为被视为异常行为。

近年来,随着深度学习和视频监控技术的成熟,结合深度学习技术对监控视频中的异常行为进行识别与监测成为计算机视觉领域的热点研究课题。其中,基于深度学习的异常行为检测流程一般包括特征提取、异常检测和异常定位等步骤。在特征提取阶段,利用卷积神经网络构造特征提取器,并对视频帧进行特征提取。与大多数的深度学习异常行为检测方法相比,本文提出的基于双尺度串行网络的视频异常行为检测模型(Dual-Scale Serial Network,DSS-Net)是基于区域分割和区域生成来进行异常行为定位。此外,受长短期记忆(Long Short-Term Memory,LSTM)[7]网络对时序特征的处理启发,DSS-Net 结合 LSTM 网络构造了一个用于捕获行为动态变化的解码网络,从而将异常行为检测问题转化为时间序列特征预测问题。在此基础上,DSS-Net 还引入了注意力机制,进一步强化关键特征的表达能力,提高模型对异常行为的定位。

1 相关工作

近年来,研究人员对基于监控视频的异常行为检测进行了大量研究。从特征提取的角度可将现有的异常行为检测方法划分为基于手工特征的异常行为检测和基于深度学习的异常行为检测。

基于手工特征的异常行为检测方法在模型构建中通常是提取正常行为或异常行为的关键特征[8],虽然该类方法取得了较好的识别结果,但模型性能主要依赖于高质量的代表性特征。如 Sharma 等[9]首先使用堆叠去噪自动编码器以无监督的方式提取多维特征表示,然后利用支持向量机(Support Vector Machine,SVM)模型构建分类预测模型,通过计算待测视频中目标物体的得分,给出异常行为的判定。Yu 等[10]结合外观和动作特征构建了一套基于视频的异常行为检测模型,该模型借助一组共享权重的孪生网络,从外观和动作上进行建模,并在编码器和解码器中引入了特征记忆模块,提高模型对异常行为的识别度。Ionescu 等[11]从实际跟踪数据的像素中学习速度和轨迹,并使用聚类树对轨迹进行聚类,这些轨迹被用来预测跟踪对象最可能的路径。Zaheer 等[12]采用相似的粒子来表示物体的关键部位,通过设定关键部位的特征并根据其变化来捕捉物体的运动趋势。Li 等[13]提出了一个用于视频异常事件检测的新模型,该模型首先基于预先训练的卷积神经网络提取深度特征,然后将映射到深度空间的深度特征作为支持向量机分类器的输入,构造了一种无监督的实时在线异常行为检测模型,并在实际场景和开源数据集上进行测试,验证了模型的有效性。

基于深度学习的异常行为检测方法主要借助神经网络模型直接从视频关键帧中学习特征映射,根据特征规则设计网络结构,并利用大规模数据集训练获得网络模型参数。如武光利等[14]基于全卷积神经网络提出了一种新的异常行为检测模型,该模型利用 AlexNet 提取关键帧的深度特征,将所提取的关键帧特征作为高斯分类器的输入,实现基于视频的人体异常行为检测。Zhou 等[15]针对搭乘扶梯场景下的实时监控,构建了一套基于视频监控的乘客异常行为识别模型。该模型首先使用 YOLOv3 对视频中的乘客位置进行识别,然后利用卷积网络对乘客的人体骨架点进行提取与定位,并根据骨架变化与初始定位之间的距离给出异常行为发生的概率值。Fan 等[16]从空间和时序维度并行地捕获帧内和帧间的目标异常行为特征,并提出了一种端到端的半监督异常行为检测方法,通过计算正常行为特征与异常行为特征之间的度量对异常行为进行检测和定位。肖利平等[17]提出

了一种基于深度学习的异常检测框架。该框架首先通过 CoSaMP 网络提取特征，然后利用多层 LSTM 网络对农田复杂监控网络中正在发生的异常或正常事件进行分类。

虽然上述方法在理论和实际应用中都获得了较好的性能，但是基于手工特征的异常行为检测方法的性能过度依赖手工提取的分类特征，过多的人工干涉造成时间成本开销较大，模型泛化性能有限。尽管基于深度学习的异常行为检测方法可以缓解基于手工特征的异常行为检测方法的局限性，但是现有的比类模型大多采用单一的视频帧内视觉特征或帧间上下文时序特征来设计检测模型，导致无法充分利用信息，进而影响模型检测的性能。为此，本研究提出了一种基于双尺度串行网络的视频异常行为检测模型（DSS-Net），通过串行提取视频帧内空间维度特征和帧间上下文时序特征，并根据特征比对结果，快速识别给定视频中的异常行为。

2 异常行为检测模型

2.1 模型结构

图 1 给出了 DSS-Net 的结构。首先，DSS-Net 使用深度可分离卷积（Depth-wise Separable Convolution，DSC）网络[18]对 Vgg-16 网络[19]进行改进，并作为视频中帧内特征提取器；接着，利用注意力机制强化帧内空间特征的表达能力，聚焦目标的强特征；然后，利用 LSTM 网络提取时间序列维度中的帧间时序特征；最后，根据输出特征与实际特征的距离偏差大小快速过滤异常特征，并用上采样操作来确定异常行为的位置。

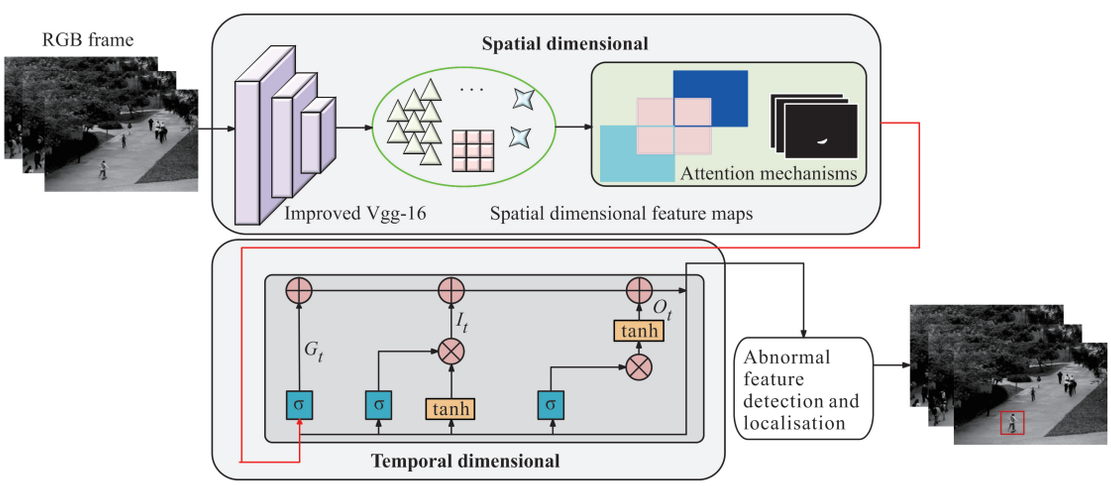


图 1 基于双尺度串行网络的视频异常行为检测模型（DSS-Net）的结构

Fig. 1 Structure of the video anomaly detection model based on dual-scale serial network (DSS-Net)

2.2 空间维度的特征提取

异常行为通常表现为不规则的形状、运动或两者的组合。由于单帧不涉及物体的运动信息和运动规律，并且异常行为检测数据集多为 2－3 s 的短视频，每秒包含 20－25 帧[2]，因此，此处选择 20 帧的帧序列组，即每间隔 2－3 s 选取 1 帧。这既可以保持动作的连续性，又可以降低模型的计算量。具体地，每一组帧序列可表示为 $S_t = < I_{t-20}, I_{t-19}, \cdots, I_{t-1} >$，其中，$I_t \in R^{w \times h}$ 表示视频中的第 t 帧。然后，利用卷积神经网络提取视频中每一帧的空间结构特征，强化异常行为和正常行为特征的区分能力。此处，以 Vgg-16 网络作为特征提取器，其结构如图 2 所示。

传统的 Vgg-16、ResNet 等特征提取网络由于结构复杂，模型参数量较多，导致模型运行时间开销较大。然而，在线异常行为的监控对于实时性要求较高，为此，本文采用 DSC 网络对传统的特征提取网络 Vgg-16 进行改进，并将改进后的 Vgg-16 网络在 ImageNet 数据集上进行预训练。DSC 网络通过将标准卷积分解为等效的多个深度卷积（Depthwise Convolution，DC）和逐点卷积来降低模型参数数量，在异常行为识别性能变化可接受的范围内尽可能地减少模型参数量，以达到降低模型的训练时间开销的目的。DSC 网络结构如图 3 所示。

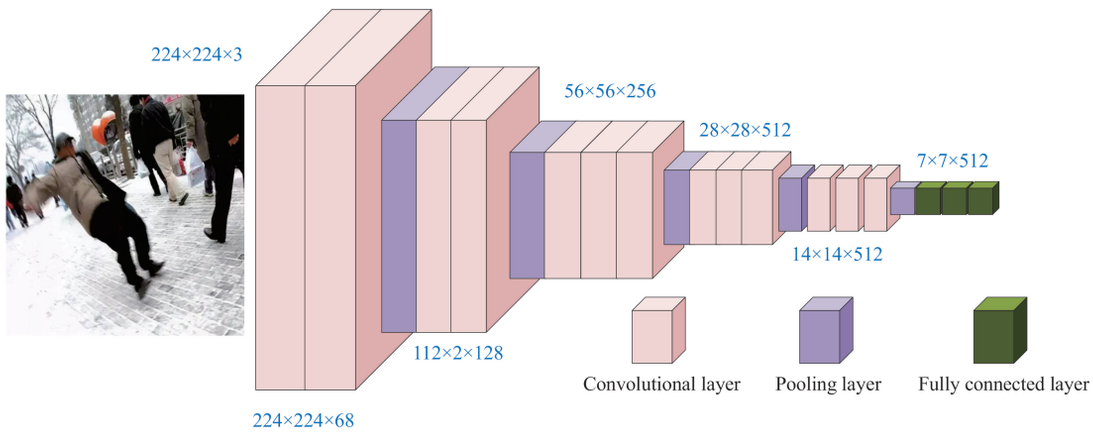


图 2　Vgg-16 网络结构

Fig. 2　Vgg-16 network structure

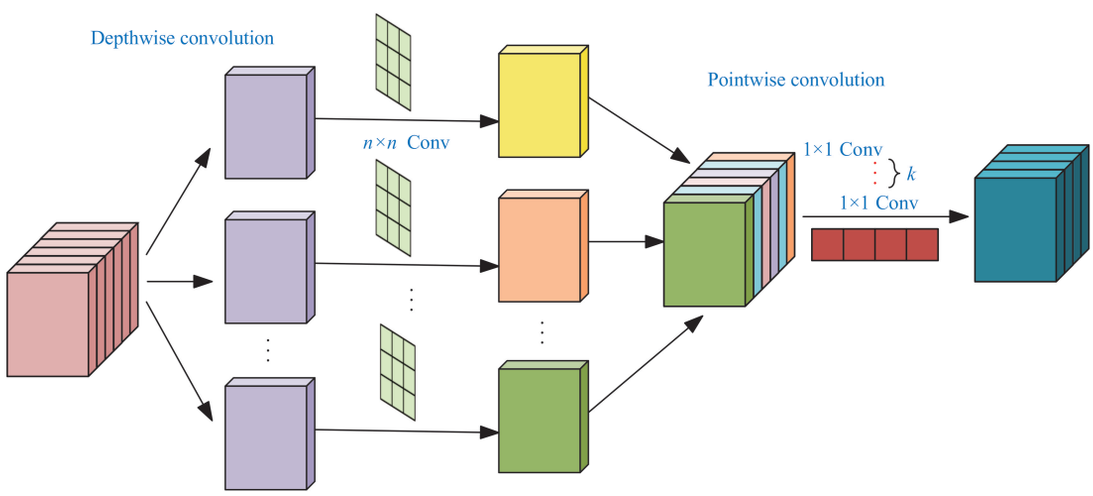


图 3　深度可分离卷积网络结构

Fig. 3　Network structure of depth-wise separable convolution

在特征提取阶段,利用图 3 所示的深度卷积将每帧 $I_t \in R^{w\times h}$ 进行卷积操作,将原始帧序列映射到深度特征空间,生成多张特征图 F_t,并利用 N 个大小为 $C_k \times C_k$ 的卷积核将 F_t 沿通道方向进行加权组合。在逐点卷积运算过程中,将 F_t 利用 k 个大小为 1×1 的卷积核进行滤波操作。上述过程简化了标准卷积的计算流程,其中,标准卷积运算过程中的参数量 $P_{\text{Vgg-16}}$ 计算如公式(1)所示:

$$P_{\text{Vgg-16}} = C_k \times C_k \times M \times N, \tag{1}$$

式中,C_k 表示卷积核的大小,M 表示通道大小。

DSC 网络在卷积操作中的参数量 P_{DSC} 计算如公式(2)所示:

$$P_{\text{DSC}} = C_k \times C_k \times M + M \times N \times 1 \times 1。 \tag{2}$$

相比标准卷积运算的参数量 $P_{\text{Vgg-16}}$,DSC 网络的参数量 P_{DSC} 下降了 Δ,其计算如公式(3)所示:

$$\Delta = \frac{P_{\text{DSC}}}{P_{\text{Vgg-16}}} = \frac{C_k \times C_k \times M + M \times N \times 1 \times 1}{C_k \times C_k \times M \times N} = \frac{1}{N} + \frac{1}{C_k^2}。 \tag{3}$$

在深度网络中,当卷积核的个数 N 较大时,公式(3)中 Δ 约为 $1/C_k^2$。特别地,常见的卷积核大小为 3×3,此时 DSC 网络的参数量大约降低为标准卷积参数量的 1/9。

2.3　注意力机制

特征提取网络中,每一层卷积块通过关注卷积核大小邻域内的部分来捕获异常行为特征。虽然一些工作通过加深网络层来增大感受野,但是深层网络不仅极易导致部分小目标物体的特征丢失,而且极易忽略全局信息[20]。此处,采用注意力机制将卷积网络的输出特征作为注意力机制的输入,强化视频帧中异常行为特征在空间维度上的表达能力。注意力机制

的原理如图 4 所示。

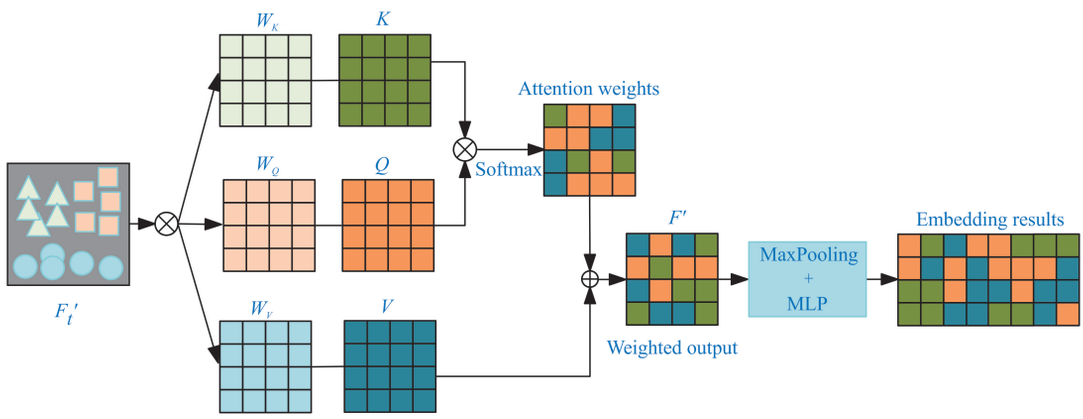


图 4　注意力机制网络结构

Fig. 4　Network structure of attention mechanism

图 4 中，首先，利用注意力机制获取帧级全局像素点和当前像素点的关系权值。计算如公式(4)和(5)所示：

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V, \quad (4)$$

$$Q=F_t'W_Q, K=F_t'W_K, V=F_t'W_V, \quad (5)$$

其中，W_K，W_Q，W_V 为可学习参数，T 表示转置操作，F_t' 表示 DSC 网络提取的空间维度特征；Q、K、V 的初始值均为 F_t'，表示 DSC 网络提取的空间维度特征，用于计算空间维度特征内部的注意力；d_k 表示特征维度。

然后，将带权重的特征 F_t' 经过最大池化层和多层感知机，获得帧级特征中的强特征。最大池化层具有旋转、平移和尺度不变性的优势，这对于视频帧中小目标或遮盖目标具有很好的适应能力，即在保留关键特征的同时防止模型过拟合。

2.4　时序维度的特征提取

视频中的异常行为通常由运动引起，主要体现在视频帧间的上下文时序信息变化中。因此，通过捕获帧间的上下文时序信息，可以有效捕获异常行为的关键特征[5]。为此，本文采用 LSTM 网络提取视频帧间的时间序列特征。此外，LSTM 网络也可以较好地缓解循环神经网络长距离依赖性差和梯度爆炸的问题。

LSTM 网络包括输入门、输出门和遗忘门。这种门结构的设计使得网络有选择性地输出关键信息，过滤边缘或无用信息，从而达到提升网络记忆能力的作用。其中，输入门决定何时读入数据到细胞单元；遗忘门决定何时忽略无用或边缘信息；输出门决定何时输出计算结果。LSTM 网络结构如图 5 所示。此处采用 LSTM 网络作为时间序列维度上的特征提取器。

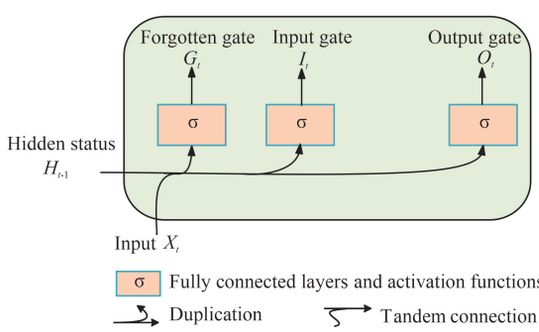


图 5　长短期记忆网络结构

Fig. 5　Long short-term memory network structure

假设在 t 时刻，空间维度特征经过注意力机制后的带权重特征图输入为 $F'a_t$，隐藏层状态为 H，上一时刻的隐藏状态为 H_{t-1}，输入门为 I_t，遗忘门为 G_t，输出门为 O_t，不同单元的计算如式(6)、(7)和(8)所示。

$$I_t=\sigma(F'a_tW_x+H_{t-1}W_{hx}+b_x), \quad (6)$$

$$G_t=\sigma(F'a_tW_y+H_{t-1}W_{hy}+b_y), \quad (7)$$

$$O_t=\sigma(X_tW_z+H_{t-1}W_{hz}+b_z), \quad (8)$$

其中，W_x，W_y，$W_z\in R^{d\times h}$ 和 W_{hx}，W_{hy}，$W_{hz}\in R^{h\times h}$ 为权重，b_x，b_y，$b_z\in R^{1\times h}$ 为偏置。利用门控单位之间的信息协作得到最终的时序维度特征。具体计算如公式(9)所示：

$$Fo_t=O_t\times\tanh(I_t\times G_{t-1}+b_t), \quad (9)$$

其中，Fo_t 表示时序维度特征，携带视频帧间的上下文信息；b_t 为偏置参数；$\times$ 表示矩阵乘法。

2.5 异常特征检测和定位

通过计算本文模型预测的特征 N_t' 和实际特征 N_t 之间的 L_2 距离,快速定位异常行为。如果距离大于设定的阈值 α,则该特征被认为是异常特征,反之则认为是正常行为。L_2 距离的计算方法如公式(10)所示:

$$L_2=\begin{cases}\text{Abnormal,if}\parallel N_t'-N_t\parallel^2>\alpha\\ \text{Normal,else}\end{cases}。\quad(10)$$

在异常特征识别并定位后,通过上采样操作反转到原始视频帧中,定位异常目标的异常行为。

3 验证实验

3.1 实验环境与评价指标

DSS-Net 的实验环境为内存 128 G 的 Ubuntu 20.4,显卡为 NVIDIA A100 GPU,显存为 40 G。采用支持 GPU 加速的 PyTorch 进行训练,cuda 环境为 NVIDIA CUDA 11.2 和 cuDNN V8.2 的深度学习加速库。模型参数的设定如表 1 所示。

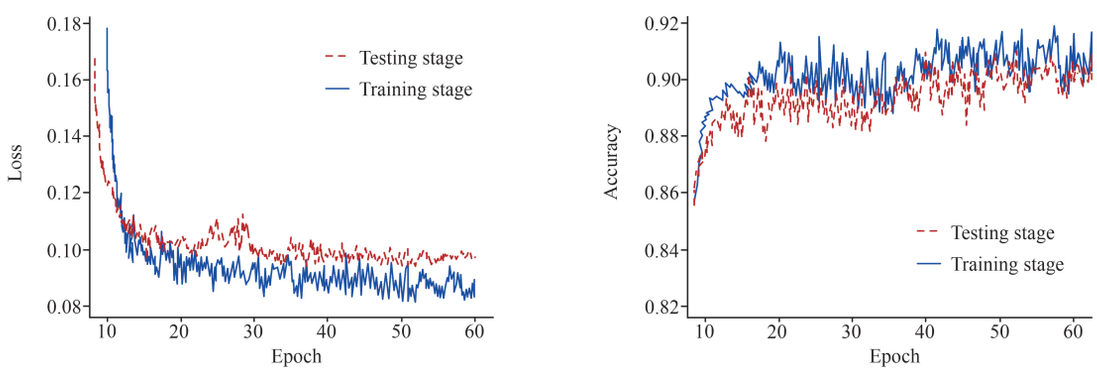
DSS-Net 训练损失曲线和准确率曲线如图 6 所示,可以看出,当模型迭代次数 Epoch 为 40 时,训练集和测试集损失曲线均趋于平稳,损失值低于 0.10,表明模型已收敛。

表 1 参数设定

Table 1 Parameter settings

参数 Parameter	取值 Value
Gradient function	SGD
Initial learning rate	0.000 5
Learning decay rate	0.001
Dropout	0.5
Batchsize	8
C_k	3
Number of 1×1 convolution kernels k	68,128,256,512
Momentum coefficient	0.9

为验证 DSS-Net 的有效性,采用识别准确率 Accuracy、等错误率(Equal Error Rate,EER)和 ROC(Receiver Operating Characteristic)线下面积(Area Under Curve,AUC)作为评价指标。ROC 曲线图中越靠近左上角,AUC 值越大,表明分类效果越好。EER 表示 ROC 曲线上与 45°角直线相交点的值,值越小,表明性能越好。



(a) Loss curves (b) Accuracy curves

图 6 DSS-Net 训练损失曲线和准确率曲线

Fig. 6 Training loss and accuracy curves of DSS-Net

3.2 实验数据集

为了验证 DSS-Net 的有效性,在主流的 UCSD[21]、UCF[22]、UMN[23]、ShanghaiTech[24] 和 CUHK Avenue[25] 等数据集上进行实验。其中,UCSD 数据集包括 Ped1 和 Ped2 两个子集,行人和行走定义为正常行为,其余如奔跑、推车、滑轮等定义为异常行为。详细信息如表 2 所示。

表 2 UCSD 数据集

Table 2 UCSD dataset

数据集 Datasets	总视频 Total video	训练视频 Training video	测试视频 Test video	帧数 Frames
Ped1	70	34	36	200
Ped2	28	16	12	120/160/180

UCF 数据集包括 1 900 个真实监控视频，950 个异常行为视频。其中，异常行为包含以下 13 种，为了便于理解，用 1－13 的数值表示每种异常行为：虐待(1)、逮捕(2)、纵火(3)、袭击(4)、事故(5)、入室盗窃(6)、爆炸(7)、打架(8)、抢劫(9)、射击(10)、偷窃(11)，入店行窃(12)、破坏公物(13)。

UMN 数据集由 3 300 帧的训练集和 4 439 帧的测试集组成，包含室内和室外两种不同场景，11 个视频片段。异常行为包括人群分散跑动和单人跑动两大类。

ShanghaiTech 数据集由 330 个训练视频片段和 107 个测试视频片段组成，包含 130 种异常行为，覆盖 13 种不同场景。

CUHK Avenue 数据集包含奔跑、抛物等 47 种不同的异常行为，由 16 个训练视频片段和 21 个测试视频片段组成。

3.3 结果与分析

3.3.1 UCSD 数据集

为验证 DSS-Net 设计的有效性，在相同的实验环境和评价指标下，分别与当前主流的视频异常行为检测模型进行对比实验，详细结果如表 3 所示。

由表 3 可以看出，在 Ped1 数据集上，DSS-Net 的 AUC 和 EER 分别达到了 95.30％和 10.60％。特别地，在 AUC 评价指标下，DSS-Net 相比对比模型中最好的 FCVA[16]、ONN[27] 以及 AOE[28]，分别提高了 0.42％(94.90％→95.30％)、0.42％(94.90％→95.30％)和 0.74％(94.60％→95.30％)；在 EER 评价指标下，DSS-Net 相比对比模型中最好的 FCVA[16] 和 SV-DNN[30]，分别降低了 6.19％(11.30％→ 10.60％) 和 33.75％ (16.00％→10.60％)。在 Ped2 数据集上，DSS-Net 的 AUC 值和 EER 值分别达到了 96.80％和 12.60％，综合优势明显。虽然在 EER 评价指标下，DSS-Net 相比 SV-DNN[30] 模型有所不足，但在 AUC 评价指标下，相比所有对比模型中表现最好的两个模型 AOE[28] 和 SV-DNN[30]，DSS-Net 分别提升了 0.94％(95.90％→ 96.80％) 和 2.00％ (94.90％ →96.80％)。究其原因是上述对比模型大多从单一的帧内空间维度或帧间上下文时序关系中推断异常行为，而所提出模型结合两方面信息，并采用串行的思想，在充分挖掘空间和时序维度信息的同时，尽可能地保留了空间和时序信息的连续性。

表 3 不同模型在 UCSD 数据集上的对比结果

Table 3 Comparison results of different models on the UCSD

数据集 Datasets	模型 Models	AUC/％ ↑	EER/％ ↓
Ped1	FCVA[16]	94.90	11.30
	HmF[26]	57.60	23.50
	ONN[27]	94.90	
	AOE[28]	94.60	
	One-SVM[29]	88.60	16.30
	SV-DNN[30]	91.90	16.00
	sHOT[31]	51.00	21.30
	DSS-Net	**95.30**	**10.60**
Ped2	FCVA[16]	92.20	12.60
	HmF[26]	89.50	17.50
	ONN[27]	94.50	
	AOE[28]	95.90	
	One-SVM[29]	90.10	15.80
	SV-DNN[30]	94.90	**12.00**
	sHOT[31]	82.90	20.90
	DSS-Net	**96.80**	12.60

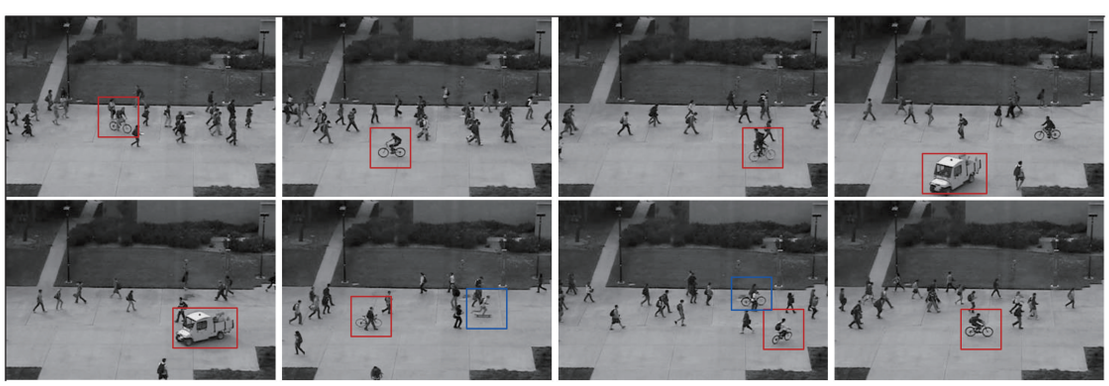
Note：the bold font represents the optimal detection result. ↑ indicates that the larger the AUC value，the better the detection result is. while ↓ indicates that the smaller the EER value，the better the detection result is.

在 UCSD Ped1 和 Ped2 数据集上的部分检测可视化结果如图 7 所示。矩形框所包围的部分为检测到的异常行为。由图 7 可以看出，DSS-Net 能够识别出滑板车、自行车、货车和奔跑等异常行为，并且 DSS-Net 对于同一帧中的不同异常行为均可以准确定位。

3.3.2 UCF 数据集

为了进一步验证 DSS-Net 的检测性能，在更具有挑战性的 UCF 数据集上进行测试。具体实验结果如表 4 所示，可以看出，DSS-Net 的 AUC 值和 EER 值分别达到了 80.60％和 18.50％。虽然在 AUC 评价指标下 DSS-Net 不及 SV-DNN[30]，但在 EER 评价指标下，DSS-Net 下降了 11.90％(21.00％→18.50％)，可以媲美当前主流的视频异常行为检测模型。

(a) Ped1



(b) Ped2

图 7　DSS-Net 在 Ped1 和 Ped2 数据集上的检测结果

Fig. 7　Detection results on Ped1 and Ped2 datasets of DSS-Net

表 4　不同模型在 UCF 数据集上的对比结果

Table 4　Comparison results of different models on the UCF dataset

模型 Model	AUC/% ↑	EER/% ↓
Two-SRNet[13]	71.00	
SV-DNN[30]	**89.20**	21.00
STFGNet[32]	80.10	
DSS-Net	80.60	**18.50**

Note: bold font represents the optimal detection result. ↑ indicates that the larger the AUC value, the better the detection result is. while ↓ indicates that the smaller the EER value, the better the detection result is.

上述数据验证了 DSS-Net 的有效性。究其原因是本文模型从时序和空间两个维度进行特征提取,这对视频帧级的上下文语义信息具有很好的捕获性。此外,DSS-Net 借助注意力机制进一步强化了空间维度上的特征映射,使得模型对类间具有更好的区分能力。

DSS-Net 在 UCF 数据集中对大部分异常行为的准确率均在 78%以上(图 8)。然而,对于袭击(4)和打架(8)行为的识别性能不高,主要原因是两者之间的行为极其相似,容易导致误报或漏报。

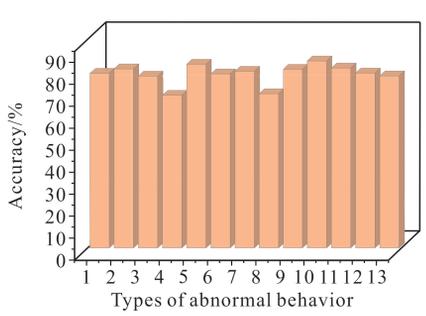


图 8　DSS-Net 在 UCF 数据集中对各类异常行为的检测准确率

Fig. 8　Detection accuracy of various anomalous behaviors of DSS-Net on the UCF dataset

为了测试 DSS-Net 的时间开销,分别与 OC-SVM、KNFST 和 HmF 模型进行对比,上述模型的

数据来源于文献[26]。对比结果如图 9 所示，可以看出，在 UCF 数据集上，DSS-Net 相比所有对比模型具有更好的实时性。特别地，相比 HmF 模型，DSS-Net 在时间开销上降低了 9.47%（0.169→0.153）。

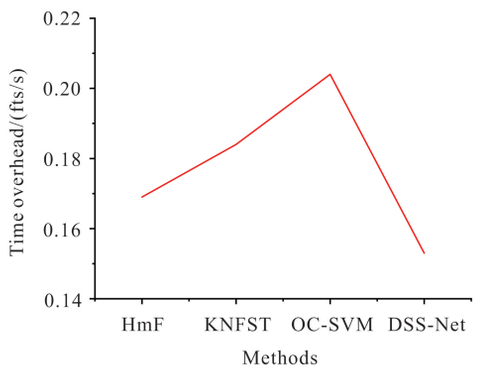


图 9　不同模型在 UCF 数据集上的时间开销对比

Fig. 9　Comparison of time consumption of different models on the UCF dataset

3.3.3　其他数据集

为了进一步验证模型的泛化性能和鲁棒性，在 UMN、ShanghaiTech 和 CUHK Avenue 等多个数据集上进行对比实验，结果如表 5 所示。

表 5　不同模型在多个数据集上的对比结果

Table 5　Comparison results of different models on multiple datasets

数据集 Datasets	模型 Models	AUC/% ↑	EER/% ↓
UMN	FCN-LSTM[14]	93.70	7.10
	Permutation[33]	91.00	
	DSS-Net	**94.61**	**5.29**
ShanghaiTech	DDGANet[34]	73.70	17.55
	MCS[35]	75.50	
	DSS-Net	**80.02**	**12.13**
CUHK Avenue	FCVA[16]	83.40	22.70
	Permutation[33]	78.30	
	MPED-RNN[36]	**86.30**	15.38
	DDGANet[34]	84.90	19.15
	DSS-Net	85.84	**14.09**

Note: bold font represents the optimal detection result. ↑ indicates that the larger the AUC value, the better the detection result is. while ↓ indicates that the smaller the EER value, the better the detection result is.

从表 5 可以看出，DSS-Net 在 3 个数据集上的检测性能相较于对比模型具有更强的竞争力。具体地，在目标更密集型的 UMN 数据集上，DSS-Net 的 AUC 值达到 94.61%，相比对比模型中表现最好的 FCN-LSTM 模型，本文模型的 AUC 值提高了 0.97%。在异常行为更加丰富的 ShanghaiTech 数据集上，DSS-Net 的 AUC 值为 80.02%，相比表现最好的 MCS 模型提高了 5.99%。虽然在 CUHK Avenue 数据集上，DSS-Net 的 AUC 值相比 MPED-RNN 有一定差距，但在 EER 指标下，DSS-Net 下降了 8.39%。究其原因是 DSS-Net 充分捕获了时序维度和空间维度上的特征，增强了特征的表达能力，有效缓解了不同异常行为之间的误报。

3.4　消融实验

3.4.1　模块间的消融实验

为了探究所提出模型中不同模块对检测性能提升的作用，接下来按照“分而治之”的机制在 UCF 数据集上进行消融实验，结果如表 6 所示。

表 6　模块间的消融实验

Table 6　Ablation experiment between modules

变体模型 Variant models	SD	DSC	AM	TD	AUC/%	EER/%	TO/(fts/s)
SD	√				76.59	26.88	1.375
SD + DSC	√	√			76.53	26.91	0.146
SD + DSC + AM	√	√	√		79.10	19.04	0.150
TD				√	68.19	29.83	0.151
SD + DSC + AM + TD	√	√	√	√	80.60	18.50	0.153

Note: SD indicates spatial dimension, AM indicates attention mechanisms, TD indicates temporal dimension, TO indicates time overhead.

由表 6 可知：①在所有模块中，空间维度特征对于模型的整体性能提升具有重要的作用，主要原因是单帧中包含的视觉信息主要来源于图片，而空间维度特征主要借助卷积神经网络进行特征提取。② DSC 可以有效改善模型的时间开销问题，使用 DSC 前后，模型的时间开销相差 8.4 倍，这进一步验证了利用 DSC 改进主干网络的合理性。③单一时序维度或空间维度特征尚不足以充分表示视频中帧内或帧间的特征，但结合两个维度的特征，在帧内视觉特征的基础上引入上下文帧间时序特征，可以有效缓解特征鲁棒性和泛化能力不强的问题。

3.4.2　阈值消融实验

为了探究阈值与 AUC 曲线之间的关系，在 UCF 数据集上进行阈值的消融实验。图 10 给出了异常特征判定的阈值与最终异常行为检测的 AUC 曲线图，可以看出，当阈值 α 为 0.06 时，AUC 值最佳。

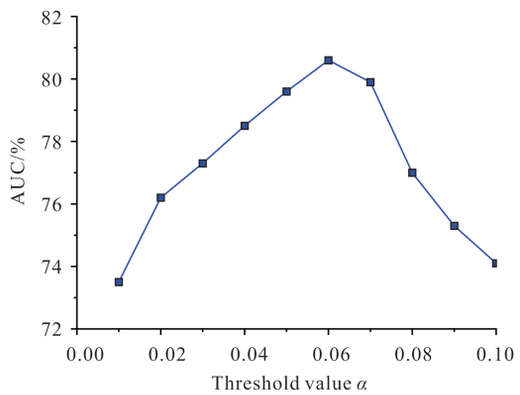


图 10　阈值 α 与 AUC 关联曲线

Fig. 10　Correlation curve between threshold value α and AUC

4　结束语

本文提出了一种新的视频异常行为检测模型。考虑到运动物体的动态变化属性,本文从时序和空间两个维度提取特征,并利用注意力机制强化目标物体的局部和全局特征,缓解特征丢失的问题。此外,考虑到时间开销问题,DSS-Net 利用 DSC 对原始主干网络 Vgg-16 进行改进,通过降低模型的参数量来减少时间开销。在 UCSD 和 UCF 两个主流数据集,以及 UMN、ShanghaiTech 和 CUHK Avenue 数据集上进行模型测试,实验结果验证了 DSS-Net 的高效性。

未来研究工作将尝试使用多尺度特征和特征融合注意力机制进一步提高 DSS-Net 的识别性能,并在实际应用中进行测试。

参考文献

[1] SREENU G,DURAI S M A. Intelligent video surveillance:a review through deep learning techniques for crowd analysis [J]. Journal of Big Data,2019,6(1):1-27.

[2] WAN S H,XU X L,WANG T,et al. An intelligent video analysis method for abnormal event detection in intelligent transportation systems [J]. IEEE Transactions on Intelligent Transportation Systems,2021,22(7):4487-4495.

[3] GEORGESCU M I,IONESCU R T,KHAN F S,et al. A background-agnostic framework with adversarial training for abnormal event detection in video [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,44(9):4505-4523.

[4] ZHOU F Q,WANG L,LI Z X,et al. Unsupervised learning approach for abnormal event detection in surveillance video by hybrid autoencoder [J]. Neural Processing Letters,2020,52(2):961-975.

[5] WANG T,QIAO M N,ZHU A C,et al. Abnormal event detection via the analysis of multi-frame optical flow information [J]. Frontiers of Computer Science,2020,14(2):304-313.

[6] 闻佳,王宏君,邓佳,等. 基于深度学习的异常事件检测[J]. 电子学报,2020,48(2):308-313.

[7] ZAYENE O,TOUJ S M,HENNEBERT J,et al. Multi-dimensional long short-term memory networks for artificial Arabic text recognition in news video [J]. IET Computer Vision,2018,12(5):710-719.

[8] BALASUNDARAM A,CHELLAPPAN C. An intelligent video analytics model for abnormal event detection in online surveillance video [J]. Journal of Real-Time Image Processing,2020,17(4):915-930.

[9] SHARMA R,SUNGHEETHA A. An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance [J]. Journal of Soft Computing Paradigm,2021,3(2):55-69.

[10] YU J,YOW K C,JEON M. Joint representation learning of appearance and motion for abnormal event detection [J]. Machine Vision and Applications,2018,29(7):1157-1170.

[11] IONESCU R T,KHAN F S,GEORGESCU M I,et al. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video [C]//2019 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach,USA:IEEE Press,2019:7842-7851.

[12] ZAHEER M Z,MAHMOOD A,SHIN H,et al. A self-reasoning framework for anomaly detection using video-level labels [J]. IEEE Signal Processing Letters,2020,27:1705-1709.

[13] LI T,CHEN X Y,ZHU F S,et al. Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection [J]. Neurocomputing,2021,439:256-270.

[14] 武光利,郭振洲,李雷霆,等. 融合 FCN 和 LSTM 的视频异常事件检测[J]. 上海交通大学学报,2021,55(5):607-614.

[15] ZHOU J T,DU J W,ZHU H Y,et al. Anomaly net:an anomaly detection network for video surveillance [J]. IEEE Transactions on Information Forensics and Security,2019,14(10):2537-2550.

[16] FAN Y X,WEN G J,LI D R,et al. Video anomaly detection and localization via gaussian mixture fully conv-

olutional variational autoencoder [J]. Computer Vision and Image Understanding，2020，195：102920.

[17] 肖利平，全腊珍，余波，等. 基于改进 CoSaMP 的农田信息异常事件检测算法[J]. 农业机械学报，2019，50(10)：230-235.

[18] ZHANG R，ZHU F，LIU J Y，et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis [J]. IEEE Transactions on Information Forensics and Security，2020，15：1138-1150.

[19] RANGARAJAN A K，PURUSHOTHAMAN R. Disease classification in eggplant using pre-trained VGG16 and MSVM [J]. Scientific Reports，2020，10(1)：1-11.

[20] ZHENG X T，ZHANG Y C，ZHENG Y P，et al. Abnormal event detection by a weakly supervised temporal attention network [J]. CAAI Transactions on Intelligence Technology，2022，7(3)：419-431.

[21] MOHAMMAD D，ALJARRAH I，JARRAH M. Searching surveillance video contents using convolutional neural network [J]. International Journal of Electrical and Computer Engineering，2021，11(2)：1656-1665.

[22] WANG T，QIAO M N，ZHU A C，et al. Abnormal event detection via covariance matrix for optical flow based feature [J]. Multimedia Tools and Applications，2018，77(13)：17375-17395.

[23] MEHRAN R，OYAMA A，SHAH M. Abnormal crowd behavior detection using social force model [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Florida，USA：IEEE Press，2009：935-942.

[24] LUO W X，LIU W，GAO S H. A revisit of sparse coding based anomaly detection in stacked RNN framework [C]//2017 Proceedings of the IEEE International Conference on Computer Vision (CVPR). Venice，Italy：IEEE Press，2017：341-349.

[25] LU C W，SHI J P，JIA J Y. Abnormal event detection at 150 fps in MATLAB [C]//2013 Proceedings of the IEEE International Conference on Computer Vision (ICCVs). Sydney，Australia：IEEE Press，2013：2720-2727.

[26] 石艳娇，张晴，崔柳，等. 基于复合运动特征的视频异常事件检测[J]. 计算机工程与设计，2020，41(10)：2872-2877.

[27] 蒋卫祥，李功. 基于一类神经网络的视频异常事件检测方法[J]. 电子测量与仪器学报，2021，35(7)：60-65.

[28] SINGH K，RAJORA S，VISHWAKARMA D K，et al. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets [J]. Neurocomputing，2020，371：188-198.

[29] 胡海洋，张力，李忠金. 融合自编码器和 one-class SVM 的异常事件检测[J]. 中国图象图形学报，2020，25(12)：2614-2629.

[30] 卢博文. 基于深度学习的监控视频中的异常行为的检测算法研究[D]. 南京：南京邮电大学，2020.

[31] RABIEE H，MOUSAVI H，NABI M，et al. Detection and localization of crowd behavior using a novel tracklet-based model [J]. International Journal of Machine Learning and Cybernetics，2018，9(12)：1999-2010.

[32] 周航，詹永照，毛启容. 基于时空融合图网络学习的视频异常事件检测[J]. 计算机研究与发展，2021，58(1)：48-59.

[33] GIORNO A D，BAGNELL A J，HEBERT M. A discriminative framework for anomaly detection in large videos[C]//2016 European Conference on Computer Vision (ECCV). Amsterdam，Netherlands：Springer International Publishing，2016：334-349.

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

[35] 张红民，庄旭，郑敬添，等. 优化 YOLO 网络的人体异常行为检测方法[J]. 计算机工程与应用，2023，59(7)：242-249.

[36] MORAIS R，LE V，TRAN T，et al. Learning regularity in skeleton trajectories for anomaly detection in videos [C]//2019 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach，USA：IEEE Press，2019：11996-12004.

Video Abnormal Behavior Detection with Dual-Scale Serial Network

WU Degang[1],ZHAO Liping[2**],CHEN Qianhui[1],ZHANG Yubo[3]
(1. College of Mechanical Engineering,Shangqiu Institute of Technology,Shangqiu,Henan,476000,China;2. College of Information and Electronic Engineering,Shangqiu Institute of Technology,Shangqiu,Henan,476000,China;3. College of Electrical and Information Engineering,Zhengzhou University,Zhengzhou,Henan,450000,China)

Abstract:Aiming at the problems of poor performance and large time overhead in traditional video abnormal behavior detection models,a video abnormal behavior detection model (Dual-Scale Serial Network,DSS-Net) based on dual-scale serial network is constructed from spatial and temporal dimensions. Firstly,the Vgg-16 network is improved by using deep separable convolution,and the improved feature extractor is used to extract features from the spatial dimension,so that the time overhead of the model can be reduced by reducing the number of calculation parameters. Then,on this basis,the attention mechanism is introduced to strengthen the expression ability of the target features. Finally,the Long Short-Term Memory (LSTM) network is used to extract the context temporal relationship between each frame of motion video from the temporal dimension. The test was carried out on the current mainstream UCSD Ped1 and Ped2 datasets and the more challenging UCF dataset. The results show that the Receiver Operating Characteristic (ROC) Area Under Curve (AUC) values of the proposed model on the three data sets reach 95.30%,96.80% and 80.60% respectively,and the Equal Error Rate (EER) reaches 10.60%,12.60% and 18.50%,respectively,and it has stronger real-time performance. Compared with the classical One-class Neural Network (ONN) and Aggregation of Ensembles (AOE) models,the AUC values of the proposed model DSS-Net are increased by 0.42% and 0.94% on the Ped1 and Ped2 datasets,respectively. In addition,the proposed model DSS-Net is also tested for generalization ability and robustness on data sets such as UMN,ShanghaiTech,and CUHK Avenue,and the results are also competitive compared with the current mainstream models.

Key words:video abnormal behavior detection;spatial dimensions;temporal dimensions;deeply separable convolution;attention mechanism

责任编辑:陆　雁