基于 Transformer 架构的端到端视频异常检测方法

李石峰*,罗　晰,刘晓茹,田　野
(渤海大学 信息科学与技术学院,辽宁 锦州 121000)

摘　要:传统的卷积神经网络虽然能够处理空间结构数据,但在处理大规模视频数据时,其时空建模能力不足。为了解决这一问题,需要一个能够处理海量视频数据的高效模型。该文提出了一种新的基于 Transformer 架构的端到端视频异常检测方法。该方法结合 Swin Transformer 架构和 Video Vision Transformer (ViViT)模型设计了时空信息融合模型,以提取视频帧序列的丰富时空信息。此外,通过将时空信息融合模型和深度支持向量数据描述(Deep SVDD)方法进行联合训练,实现了端到端的视频异常检测。在两个公开视频数据集上与最新的 10 种方法进行了对比实验,在 UCSD Ped2 数据集上,该模型取得了最高的 96.5% 的 AUC;在 CHUK Avenue 数据集上,该模型也取得了 80.7% 的 AUC,优于多数方法。与领先的视频异常检测方法相比,该方法具有一定的优势和竞争力。

关键词:视频异常检测;Transformer 架构;时空信息融合模型;深度支持向量数据描述;联合训练
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2025)06-0049-07
doi:10.20165/j.cnki.ISSN1673-629X.2025.0018

An End-to-end Video Anomaly Detection Method Based on Transformer Architecture

LI Shi-feng*,LUO Xi,LIU Xiao-ru,TIAN Ye
(School of Information Science and Technology,Bohai University,Jinzhou 121000,China)

Abstract:Although the traditional convolutional neural network can process spatial structure data,its spatiotemporal modeling ability is insufficient when processing large-scale video data. In order to solve this problem,an efficient model that can handle massive video data is needed. A new end-to-end video anomaly detection method based on Transformer architecture is proposed. Combining Swin Transformer architecture and Video Vision Transformer (ViViT) model,a spatio-temporal information fusion model is designed to extract rich spatio-temporal information of video frame sequences. In addition,by combining spatiotemporal information fusion model and Deep SVDD method,end-to-end video anomaly detection is realized. A comparison experiment was conducted on two public video datasets with the latest 10 methods. On UCSD Ped2 dataset,the proposed model achieved the highest AUC of 96.5%. On the CHUK Avenue dataset,it also achieved 80.7% AUC,which is better than that of most methods. The proposed method has certain advantages and competitiveness compared with the leading video anomaly detection methods.

Key words:video anomaly detection;Transformer architecture;spatio-temporal information fusion model;deep support vector data description (Deep SVDD);joint training

0　引　言

随着人们的安全意识逐渐提高,越来越多的监控摄像头出现在日常生活中。及时、准确地检测视频异常对于确保公共安全和维持社会秩序至关重要。但由于环境条件的复杂性和人类行为的不确定性,视频中的异常具有模糊性、稀缺性,并且正、负样本极度不平衡,使得视频异常检测成为具有挑战性的任务。

视频异常检测首先需要确定异常的含义,一般来说在不同的视频中对于“异常”的定义各不相同,即异常的定义取决于视频本身的内容,通常情况下将视频场景中小概率事件视为异常行为[1]。Chandola 等人[2]认为视频中的异常检测是指识别不符合预期行为的事件。

由于视频集成了时间和空间信息,因此视频异常检测网络模型需要能同时提取时空特征。目前多数文献都是基于深度学习的方法,如三维卷积神经网络

收稿日期:2024-10-12　　**修回日期**:2025-02-16
基金项目:国家自然科学基金(61402049);辽宁省教育厅科学研究项目(LJ212410167033);辽宁省社会科学规划基金(L21BGL002)
作者简介:李石峰(1980-),男,副教授,硕导,博士,通讯作者,研究方向为计算机视觉;罗　晰(2001-),女,硕士研究生,研究方向为计算机视觉。

(3D CNN)、自编码器(AE)或者生成对抗网络(GAN)等进行视频异常检测建模。鉴于 Transformer 在自然语言处理领域取得了巨大的成功,越来越多的研究人员开始将 Transformer 应用于计算机视觉领域。例如,ViT[3]将 16 × 16 个图像补丁作为输入到 Transformer 编码器,实现图像分类。ViViT[4]基于 ViT,探索了 ViT 在视频分类中的应用。

基于 Transformer 在序列建模方面的惊人表现,受 ViViT 时空序列建模方式的启发,该文研究如何使用 Transformer 架构获取视频序列的时空信息,并设计了时空信息融合模型。主要贡献如下:

(1)鉴于 Transformer 在时空序列方面强大的建模能力,提出了一种基于 Transformer 的视频异常事件检测,该框架能够实现端到端视频异常检测。

(2)设计了一种基于纯 Transformer 结构的时空信息融合模型,并且有效融合 Deep SVDD 方法,实现了端到端的异常检测。

1 相关工作

1.1 视频异常检测方法研究

在视频异常检测领域,深度学习成为目前较为主流的研究方法。而当前比较广泛采用的建模方法主要有以下几类。

(1)基于分类的异常检测。基于分类的异常检测方法将视频中的正常和异常行为视为两个类别,利用二分类思想进行视频异常检测。Deepak 等人[5]通过混合表示学习和双流深度特征提取技术,从不同视图提取特征,利用视图间的互补信息全面表示数据,最后通过 One-class SVM(OC-SVM)区分正常与异常事件。胡学敏等人[6]通过光流法计算等间距采样的特征点光流场,将图像划分为子区域并计算时空立方体特征;设计基于最近邻分类和支持向量机的级联分类器,完成人群异常行为的检测与定位。胡薰尹等人[7]通过输入对光照和背景变化不敏感的三通道矫正光流运动历史图至 3D 卷积核提取短时序特征,结合可学习贡献因子的 3D-LCRN 网络对 COFMHI 进行分类识别,提取多层次时空特征。Gong 等人[8]将 YOLO 提取的前景人体作为 3D CNN 的输入提取行为的时空特征进而分类正、异常行为。

(2)基于重构误差的建模。基于重构误差的建模思想是建立在假设异常事件会产生较大的重建误差基础上从而实现视频异常检测。卷积自编码器常用于重构输入图像,Zhao 等人[9]通过三维卷积从空间和时间维度提取图像特征,结合重建损失和权重递减的预测损失生成未来帧,增强运动特征学习。Ribeiro 等人[10]使用卷积自编码器(CAE)对原图,边缘图以及光流图的重构误差作为异常评分。然而二维卷积无法捕获时序信息。Ravanbakhsh 等人[11-12]使用 GAN 网络训练正常帧及其光流图像,只用正常数据训练使 GAN 无法生成异常事件,计算真实数据与重建的外观和运动表示的重构误差,通过局部差异检测异常区域。岳海纯[13]采用 3D 编码+LSTM+3D 反卷积解码的网络结构对图像进行重构,通过在编码器和解码器之间增加跳越层连接,使得重构图像更加完整。Chang 等人[14]设计了一种新型卷积自动编码器架构,分别捕获时空信息,空间部分重构最后一帧,时间部分以连续帧输入、RGB 差值输出,模拟光流产生。异常事件会导致较大重构误差。Zaheer 等人[15]将鉴别器的作用从识别真实与虚假数据转为区分良好与差质量的重建,进行广泛训练产生稳定结果,实现生成器与鉴别器的对抗性模型高效、鲁棒地进行异常检测。

(3)基于预测的模型。该类模型假设正常行为是有规律且可预测的,而异常行为是不可预测的,通过预测误差实现检测异常行为。Liu 等人[16]使用 U-Net 作为预测网络,除了对抗性训练和外观约束外,首次在模型中加入时间约束,通过从外观和运动两方面预测正常事件。Li 等人[17]结合 U-Net 在表示空间信息方面的优点和 Conv-LSTM 建模时间运动数据的能力,设计了未来帧的预测模型。Chen 等人[18]基于 U-Net 提出了双向预测模型,通过正向和反向子网预测帧与目标帧的相似性,在测试阶段采用滑动窗口方案聚焦视频帧中的异常区域。Dong 等人[19]提出了一种基于双鉴别器的 GAN 半监督方法,框架由生成器和两个分别区分外观与运动真假的鉴别器组成,通过交替优化生成器和鉴别器提升预测性能。Li 等人[20]提出了一种解耦架构来学习时空信息,模型分为两部分:第一部分从单帧视频中提取外观特征,第二部分利用这些特征预测未来帧的潜在代码,通过预测和观测潜在码差异作为异常度量。Huang 等人[21]通过设计双流编码器编码外观和运动信息,引入约束提高其特征语义一致性,正常样本保持高一致性,而异常样本的外观与运动特征一致性较低,导致较大预测误差,便于异常检测。

1.2 基于 Self-Attention 的 Transformer 模型

Transformer 是第一个完全依赖自注意力(Self-Attention)来计算输入和输出的模型架构,而没有使用结构对齐的 RNN 或 CNN。Transformer 模型结构包含两个部分:编码器(encoder)和解码器(decoder)。编码器由 6 个相同的层组成,每层有两个子层:多头注意力机制(Multi-Head Attention,MHA)和全连接前馈网络(Feed Forward,FF),并采用残差(residual)连接和层归一化。解码器结构与编码器类似,但额外添加了

一个子层,对编码器输出进行掩码(Masked)操作。

1.3　Swin Transformer 模型

为解决 Transformer 在 NLP 领域的高性能迁移到计算机视觉任务的问题挑战,Liu 等人[22]提出了一种新的视觉 Transformer,称为 Swin Transformer。通过将 Self-Attention 计算限制在非重叠局部窗口并允许跨窗口连接,提升了计算效率。其层次结构在不同尺度上建模,具有与图像大小相关的线性计算复杂性,显著超越了之前的技术水平,展示了 Transformer 在视觉任务中的潜力。

1.4　ViViT 模型

借鉴了 ViT 在图像分类方面的最新成果,Arnab 等人[4]提出了一种基于 Transformer 的视频分类模型 ViViT。通过提取时空 tokens 并将 Transformer 编码器分解为空间和时间部分,分别处理空间和时间信息。尽管基于 Transformer 的模型通常需要大型数据集,ViViT 展示了如何通过规范化模型和利用预训练图片模型在小数据集上有效训练。该模型在多个视频分类基准数据集上取得了最先进的结果,优于之前的深度 3D 卷积方法。

2　文中方法

2.1　数据预处理

由于深度学习中数据量巨大,前向计算和反向传播难以一次完成,通常通过顺序打乱数据并分割成小块进行预处理,确保数据完整性。然而,这样既耗时又可能造成损失。为解决这一问题,该文提出了一种基于多特征融合的方法,快速处理视频数据并将其转化为符合模型要求的输入形式,以提高异常检测的准确性。

假定训练和测试数据集中含有 N 个视频,视频 V_i ($i=1,2,\cdots,N$)被均匀剪切成多个连续帧,这些视频帧具有时间连续性。针对每一段视频,为了确保模型提取特征时的时间连续性得到充分考虑,该文采取多特征融合的方法来获取模型所需的输入 X,从而达到更高层次的表达:从整段视频帧序列的第一帧开始,取连续的 t 帧作为第一个输入 x_1,然后每隔 $t/2$ 帧进行下一个 x 的获取。假设视频 V_i 的帧序列长度为 L,则每段视频可获取的 x 的数量为 $M=(L-t)/(t/2)+1$。在此基础上,对 M 的数值进行向下取整,即每段视频最后不满足 t 帧的序列会被舍弃,最终每段视频 V_i 生成的所有 M 个序列 $X=\{x_1,x_2,\cdots,x_M\}$ 会作为时空信息融合模型的输入。

2.2　整体模型架构

设计的模型总体结构如图 1 所示。

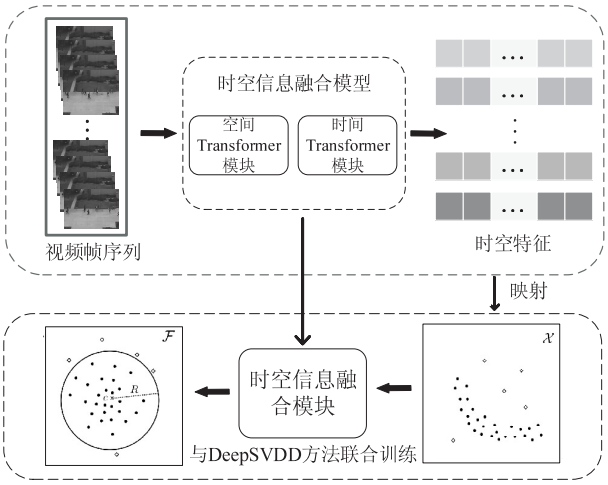

图 1　模型整体架构

整体框架由两部分组成,上半部分为基于纯 Transformer 架构设计的时空信息融合模型,下半部分为时空融合模型和 Deep SVDD 方法共同训练实现端到端视频异常检测。其中,时空融合模型由空间 Transformer 模块与时间 Transformer 模块组成。

2.3　时空信息融合模型

空间 Transformer 模块(spatial transformer module)用于对输入的视频序列进行空间信息编码,以获取融合空间信息的特征张量。基于 Swin Transformer 在 CV 任务中的优异表现,该文对其进行微调和结构修改,将其应用于时空信息融合模型作为空间特征提取器。如图 2 所示,迁移 Swin Transformer 在 ImageNet 上的预训练模型,去除预测头并添加自适应平均池化层(AdaptiveAvgPool1d)和层归一化层(LN),旨在压缩输出特征,消除冗余信息,降低参数量,训练中对模型部分参数进行冻结。

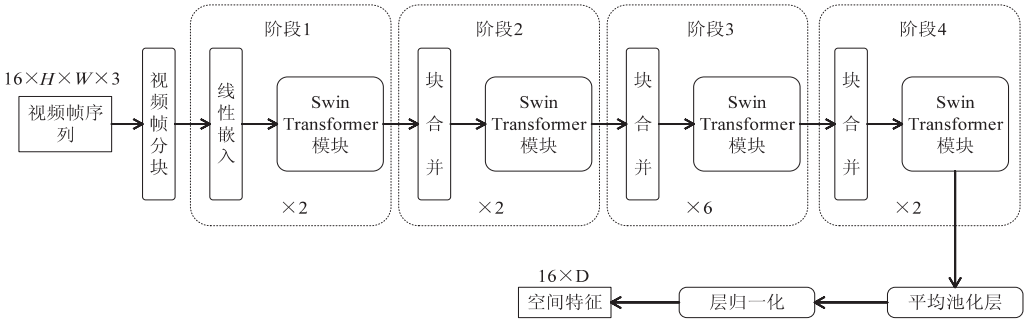

图 2　空间 Transformer 模块

时间 Transformer 模块(temporal transformer module)用于处理视频异常检测中的时序信息,强调时序信息与空间信息同样重要。以往通过 FlowNet 提取时间信息,但 FlowNet 在检测框架中会显著增加模型参数。受 ViViT 启发,该文设计了基于 MHA 机制的时间 Transformer 模块提取时序信息。如图 3 所示,改进的 Transformer Encoder 作为时序信息提取器,LN 层被添加在 MHA 和 FF 层之前进行层归一化。

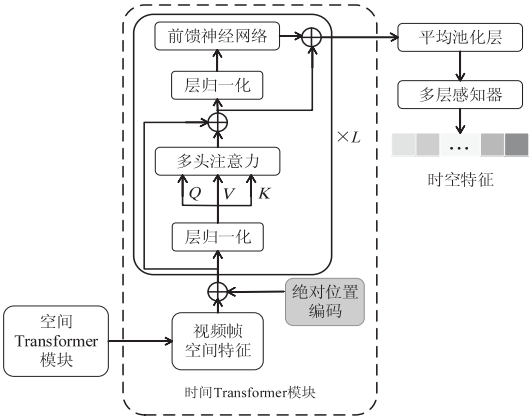

图 3 时间 Transformer 模块

每个 MHA 和前馈神经网络层后也都添加了残差连接,避免梯度消失,同时降低模型复杂度,从而减少过拟合。最后经过一个自适应平均池化层和一个多层感知器层处理后,时空信息融合模型的输出是融合了输入视频帧序列时空信息的特征张量。

绝对位置编码(absolute position coding)。位置编码对 Transformer 结构具有重要的影响,该文在时空信息融合模型架构中增加了时序绝对位置编码嵌入(如图 3 灰色区域所示)。传统 Transformer 架构位置编码采用三角函数式绝对位置编码结构,该文则是利用训练式绝对位置编码结构来获取视频帧序列的时间信息。该编码方式把位置编码作为可训练参数来使用,如最大长度是 1 000,编码维度是 192,则初始化 1 000×192 矩阵作为位置向量使其随训练过程不断更新。该文没有采用 Swin Transformer 中用到的相对位置编码是由于加入的位置编码仅对时序信息有效,对单个视频帧序列的时间位置信息根本无需用到繁杂的相对位置编码。实验验证了加入绝对位置嵌入后有助于改善模型的性能,具体内容见第 3 章相关消融实验。

2.4 Deep SVDD 联合训练模型

Ruff 等人[23]提出了一种深度支持向量数据描述(Deep Support Vector Data Description,Deep SVDD)的异常检测方法,该方法基于异常检测的目标进行训练,是一种单类(One-Class)分类方法。

该文旨在实现端到端的视频异常检测,故通过将 soft-boundary Deep SVDD 方法与时空信息融合模型进行联合训练设计了基于单类分类的端到端视频异常检测方法。训练过程中,Deep SVDD 对模型参数 W 进行优化的同时训练得到一个可以包含所有正常训练样本的最小体积的超球体,其中超球体的半径 $R>0$,中心点为 c。设置训练中的损失函数为:

$$\text{Loss}=R^2+\frac{1}{vn}\sum_{i=1}^{n}\max\{0,(f_{\text{out}}-c)^2-R^2\}+\frac{\lambda}{2}\sum_{l=1}^{L}\|W^l\|_F^2 \tag{1}$$

该损失函数主要包括三项,第一项为超球体的半径,模型的目的是最小化 R,使超球体的体积最小化。第二项是样本惩罚项,如果该样本通过网络后,到超球体中心的距离大于超球体半径 R,则受到惩罚。其中,f_{out} 是时空信息融合模型的输出,n 是输入样本空间所有的输入视频序列的张量总数,超参数 $v\in(0,1]$ 控制了超求球体的体积和异常样本边界之间的权衡,即允许一些点被映射到球体的外部,实验中设置 $v=0.01$,超球体中心 c 的选择必须不是全零解,并且在 Deep SVDD 中只使用无偏差项或有界激活函数的神经网络才能有效防止平凡解的产生,将 c 固定为对一些训练数据样本进行初始向前传递后产生的网络表示的平均值是一个很好的策略,该文也是采用同样的策略来初始化 c 防止平凡解的产生。优化损失函数让模型学习参数 W,使数据点更加紧密地映射到超球体内。最后一项是具有超参数 $\lambda>0$ 的网络参数 W 上的权值衰减正则化器,其中 $\|\cdot\|_F$ 表示弗罗比尼乌斯范数。

由于网络参数 W 和半径 R 通常处在不同的尺度上,使用常见的随机梯度下降 SGD 算法可能无法优化模型参数。因此,采取交替优化的方法对网络参数 W 和半径 R 进行优化。在开始训练的时候通过固定初始化的半径 R,训练 k epochs($k<<N$,N 为整个训练过程设置的最大 epoch 值)的网络模型参数 W。然后,每训练 k 个 epochs,利用最新一次更新后网络的参数 W 并由给定网络数据表示解算出半径 R 并更新。实验中,使用 AdamW 算法对参数进行优化,这种交替优化模型参数 W 和超球体半径 R 的方法,有助于更好地完成视频异常检测任务。

2.5 异常评分

训练数据集只包含正常帧,因此,在测试阶段,包含异常帧的视频序列在经过时空信息融合模型后输出 f_{out} 距离超球体中心 c 的距离会比正常样本的距离更远,即大于 R。通过计算模型输出 f_{out} 与超球体中心 c 的欧氏距离,并计算该距离与 R 的平方差值作为判断该视频序列是否包含异常的标准,异常分数计算公式如下:

$$\text{Score} = |f_{\text{out}} - c|^2 - R^2 \tag{2}$$

显然,如果 Score 的值大于 0,则说明模型输出 f_{out} 到中心 c 的距离大于 R,即当前视频序列中包含异常帧,反之 Score 的值小于或等于 0,说明视频序列中不包含异常。

3　实　验

3.1　实验环境

该文采用 Pycharm 作为程序开发 IDE 工具,Python 作为基础开发语言,Python 版本为 3.7,使用 PyTorch1.7.1 作为深度学习开发框架,所有代码在深度学习服务器上进行编译和调试。服务器 CPU 为 16 vCPU Intel(R) Xeon(R) Platinum 8350C CPU @ 2.60 GHz,GPU 为 Geforce RTX 3090。程序运行中主要使用的工具包有 Opencv、Numpy 等。实验的硬件环境如表 1 所示。

表 1　实验环境配置

运行环境	详细配置
开发语言	Python3.7
深度学习框架	PyTorch1.7
前端技术	Opencv、Numpy 等
开发工具	Pycharm
GPU 处理器	RTX 3090
CUDA 版本	10.1

3.2　数据集

实验中主要使用了 CUHK Avenue、UCSD Ped2 数据集。一些正异常样本示例如图 4 所示。

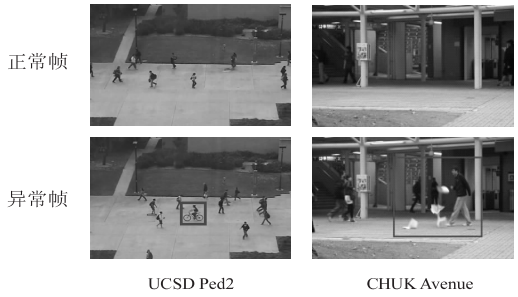

图 4　一些正异常样本

CUHK Avenue 数据集包含 16 个训练视频和 21 个测试视频,共有 47 个异常事件,包括投掷物体、闲逛和跑步。视频中人的大小可能会因为摄像头的位置和角度不同而变化。

UCSD Ped2 数据集包含 16 个训练视频和 12 个测试视频,其中有 12 个异常事件。异常事例都是关于自行车和汽车等车辆的。

3.3　评价标准

ROC(Receiver Operating Characteristic)曲线,又称接受者操作特征曲线,是通过绘制真正率(True Positive Rate, TPR)与假正率(False Positive Rate, FPR)之间的关系来构建的。AUC(Area Under Curve)被定义为 ROC 曲线下的面积,可以通过公式 3 计算。目前大多使用 AUC 值作为模型的评价标准是因为很多时候 ROC 曲线并不能清晰地说明哪个分类器的效果更好。在视频异常检测的文献中,AUC 也是被用作评价模型优劣的标准指标。AUC 检测值越高,表示异常检测性能越好。该文利用帧级 AUC 进行性能评估。

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{3}$$

3.4　实验细节

整个模型以视频帧序列作为输入,实验中设置 t = 16 帧组成一个视频序列,然后按照时间顺序每隔 $t/2$ 帧(即 8 帧)取下一个输入序列。每一帧图片大小被处理为 224×224×3,则时空信息融合模型的输入视频序列的大小为 16×224×224×3,模型的输出是一个 1×192 的潜在特征向量。该文迁移学习所采用 Swin Transformer 的预训练模型参数配置为:Swin-T: C = 96, layer numbers = {2,2,6,2},其中 C 表示通道数,layer numbers 表示每个阶段的 Swin Transformer 块的数量。训练过程中,训练 epoch = 200、batchsize = 4。针对时间 Transformer 模块中 L 的取值问题,实验中设置 L = 4 时,提出的模型方法在公开数据集上取得的效果最优。

使用 AdamW 作为参数优化器,参数设置为 β_1 = 0.9, β_2 = 0.999, ε = 1e-8。因为固定值的学习率很容易最终不断振荡,无法实现收敛,所以实验中使用了学习率衰减算法训练模型,初始学习率为 lr = 1e-4,当连续 5 次训练损失不下降时,学习率衰减 10 倍。

3.5　方法比较

这一节比较了提出的方法与现有的一些基于深度学习的方法,包括基于分类、基于重建以及基于预测等方法。不同方法的 AUC 列于表 2 中。从表中可以看到,该方法在 UCSD Ped2 数据集上取得了惊人的 96.5% 的 AUC 值,优于大多数现有的方法;因为 Transformer 结构在处理复杂背景时容易产生过拟合问题,容易降低异常检测的准确率,而 Avenue 数据集的训练数据中的背景相对于 Ped2 更加复杂以及摄像头的位置和角度等问题,在 Avenue 数据集上的实验没有取得最先进的 AUC,但仍有不错的表现。这足以说明该方法的有效性。

上述两个公共数据集上对提出模型的性能进行了基准测试,全部的训练和测试都是在 NVIDIA GeForce 架构的图像处理器上进行的,测试阶段平均运行时间约为 145 fps,如表 3 所示,这已远远超过了 25 fps[25],20 fps[31] 以及 42.5 fps[32] 的运行时间,虽然平均时间

没有超越 150 fps[33],但已经可以满足大多数视频的实时要求,在最新的研究中也具有竞争力。

表 2 Ped2 和 Avenue 数据集上不同方法的 AUC 比较 %

Method	UCSD Ped2	CUHK Avenue
AE-Conv3D[24]	91.2	77.1
Conv-AE[25]	90.0	70.2
Del Giorno et al. 2016[26]	-	78.3
Unmasking[27]	82.2	80.6
Stacked RNN[28]	92.2	81.7
Future Frame Prediction[16]	95.4	85.1
AbnormalGAN[17]	93.5	-
Memory-Augmented[29]	94.1	83.3
Dual Discriminator[19]	95.6	84.9
Multi-Task Learning[30]	92.4	86.9
Our Method	96.5	80.7

表 3 运行时间比较

Method	Running Time/fps
Liuet al. 2018[16]	25
Cai et al. 2021[31]	20
Lu et al. 2013[32]	150
Duman et al. 2019[33]	42.5
Ours	145

3.6 消融实验

与 ViViT 基准模型结果比较。受 ViViT 模型时空信息串联融合思想的启发,设计了所提方法的时空信息融合模型,通过设置同样的参数,同样将 ViViT 模型与 Deep SVDD 方法进行联合训练,在 UCSDped2 数据集以及 CUHK Avenue 数据集上做了消融实验。实验结果如表 4 所示,在两个公开数据集上,设计的模型与 ViViT 模型相比分别取得了 10 百分点以上的 AUC 值,这表明所设计的模型与 ViViT 基准模型相比可以取得更好的检测精度。

表 4 消融实验结果对比 %

	UCSD Ped2	CUHK Avenue
ViViT(base model)	82.4	67.8
Our Method	96.5	80.7
abs. pos.	96.5	80.7
rel. pos.	89.6	75.8
with pre. model	96.5	80.7

• 绝对位置编码的作用。通过添加不同的位置嵌入方式进行实验,发现在模型中加入绝对位置编码(absolute position coding)的效果要比不加位置嵌入或加入相对位置编码的实验效果更优,实验结果如表 4,其中,abs. pos. 表示添加绝对位置编码,rel. pos. 表示添加相对位置编码。

绝对位置编码相比其他编码方法在时空信息融合中表现优越,能够提供精确的时空位置信息,有助于捕捉视频中的长时间依赖关系和时空结构。其全局固定性避免了相对位置编码的动态变化和冗余问题,显著提升了异常检测的准确性和鲁棒性。因此,绝对位置编码在视频异常检测中对模型性能的提升起到了关键作用。

• 是否加载预训练模型结果比较。为了探讨加载预训练模型对所提模型对视频异常检测效果的影响,通过设置同样的参数,不加载预训练模型参数在 UCSD ped2 数据集以及 CUHK Avenue 数据集上同样做了实验,消融实验结果如表 4 所示,其中 with pre. model 表示加载预训练模型。结果表明加载预训练模型参数对模型效果的提升有很大帮助。

4 结束语

提出了一种基于 Transformer 的视频异常事件检测架构,该架构能够实现端到端视频异常检测。Transformer 在建模时空序列方面有比肩甚至更优于深层神经网络的强大能力,在抛除传统卷积的基础上,提出利用 Transformer 架构实现视频异常检测任务。该文设计了一种基于纯 Transformer 结构的时空信息融合模型,并且有效的与 Deep SVDD 方法结合进行联合训练,实现了完全端到端的异常检测。在现有的公共数据集上进行的实验表明,该视频异常检测方法是有效的,一定程度上实现了视频异常及时准确的检测。该研究工作存在一些不足之处,比如在面对复杂背景环境时未能达到最先进的检测效果,在下一步工作中,将继续基于 Transformer 架构进行模型的改进与设计,进一步优化模型方法的性能。

参考文献:

[1] 胡正平,张 乐,李淑芳,等. 视频监控系统异常目标检测与定位综述[J]. 燕山大学学报,2019,43(1):1-12.

[2] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3):1-58.

[3] DOSOVITSKIY A. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. 2020[2024-10-03]. https://arxiv.org/abs/2010.11929.

[4] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: a video vision transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. [s. l.]: IEEE, 2021:6836-6846.

[5] DEEPAK K, SRIVATHSAN G, ROSHAN S, et al. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders[J]. Circuits, Systems, and Signal Processing, 2020, 40(3):1-17.

[6] 胡学敏,余 进,邓重阳,等. 基于时空立方体的人群异常行为检测与定位[J]. 武汉大学学报:信息科学版,2019,44

(10):1530-1537.

[7] 胡薰尹,管业鹏. 基于 3D-LCRN 视频异常行为识别方法[J]. 哈尔滨工业大学学报,2019,51(11):183-193.

[8] GONG M G,ZENG H M,XIE Y,et al. Local distinguishability aggrandizing network for human anomaly detection[J]. Neural Networks,2020,122:364-373.

[9] ZHAO Y,DENG B,SHEN C,et al. Spatio-temporal autoencoder for video anomaly detection[C]//Proceedings of the 25th ACM international conference on multimedia. Mountain View:ACM,2017:1933-1941.

[10] RIBEIRO M,LAZZARETTI A E,LOPES H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters,2018,105:13-22.

[11] RAVANBAKHSH M,NABI M,SANGINETO E,et al. Abnormal event detection in videos using generative adversarial nets[C]//2017 IEEE international conference on image processing (ICIP). Beijing:IEEE,2017:1577-1581.

[12] RAVANBAKHSH M, SANGINETO E, NABI M, et al. Training adversarial discriminators for cross-channel abnormal event detection in crowds[C]//2019 IEEE winter conference on applications of computer vision (WACV). Snowbird:IEEE,2019:1896-1904.

[13] 岳海纯. 基于自动编码器的异常行为检测[D]. 长春:吉林大学,2020.

[14] CHANG Y,TU Z,XIE W,et al. Clustering driven deep autoencoder for video anomaly detection[C]//Computer vision-ECCV 2020: 16th European conference. Glasgow: Springer,2020:329-345.

[15] ZAHEER M Z,LEE J H,ASTRID M,et al. Old is gold:redefining the adversarially learned one-class classifier training paradigm[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]: IEEE, 2020:14183-14193.

[16] LIU W,LUO W,LIAN D,et al. Future frame prediction for anomaly detection—a new baseline[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:6536-6545.

[17] LI Y Y,CAI Y H,LIU J Q,et al. Spatio-temporal unity networking for video anomaly detection[J]. IEEE Access, 2019,7:172425-172432.

[18] CHEN D Y,WANG P T,YUE L Y,et al. Anomaly detection in surveillance video based on bidirectional prediction[J]. Image and Vision Computing,2020,98:103915.

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

[20] LI B,LEROUX S,SIMOENS P. Decoupled appearance and motion learning for efficient anomaly detection in surveillance video[J]. Computer Vision and Image Understanding, 2021,210:103249.

[21] HUANG X,ZHAO C,WU Z. A video anomaly detection framework based on appearance-motion semantics representation consistency[C]//ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing. Rhodes:IEEE,2023:1-5.

[22] LIU Z,LIN Y,CAO Y,et al. Swin transformer:hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. [s. l.]:IEEE,2021:10012-10022.

[23] RUFF L,VANDERMEULEN R,GOERNITZ N,et al. Deep one-class classification[C]//International conference on machine learning. Stockholmsmässan: PMLR, 2018: 4393-4402.

[24] JI S W,XU W,YANG M,et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,35(1): 221-231.

[25] HASAN M,CHOI J,NEUMANN J,et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:733-742.

[26] DEL GIORNO A,BAGNELL J A,HEBERT M. A discriminative framework for anomaly detection in large videos[C]//Computer vision-ECCV 2016:14th European conference. Amsterdam:Springer,2016:334-349.

[27] TUDOR IONESCU R,SMEUREANU S,ALEXE B,et al. Unmasking the abnormal events in video[C]//Proceedings of the IEEE international conference on computer vision. Venice:IEEE,2017:2895-2903.

[28] LUO W,LIU W,GAO S. A revisit of sparse coding based anomaly detection in stacked RNN framework[C]//Proceedings of the IEEE international conference on computer vision. Venice:IEEE,2017:341-349.

[29] GONG D,LIU L Q,LE V,et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:1705-1714.

[30] GEORGESCU M I,BARBALAU A,IONESCU R T,et al. Anomaly detection in video via self-supervised and multi-task learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]: IEEE,2021:12742-12752.

[31] CAI Y H,LIU J Q,GUO Y J,et al. Video anomaly detection with multi-scale feature and temporal information fusion[J]. Neurocomputing,2021,423:264-273.

[32] LU C,SHI J,JIA J. Abnormal event detection at 150 fps in Matlab[C]//Proceedings of the IEEE international conference on computer vision. Sydney:IEEE,2013:2720-2727.

[33] DUMAN E,ERDEM O A. Anomaly detection in videos using optical flow and convolutional autoencoder[J]. IEEE Access,2019,7:183914-183923.