文章编号：2095-2163(2021)06-0070-06 中图分类号：TP183 文献标志码：A

基于 3D 卷积自编码器的视频异常行为检测

连 靖，胡 兴，黄影平
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：视频异常行为检测是目前计算机视觉领域的热点问题之一。然而，由于异常行为难以具体定义，使得基于监督学习的二类分类方法难以应用在该领域。本文提出了一种无监督的视频异常检测模型，称之为基于时空特征融合的 3D 自编码器模型（ST-3DCAE）。模型采用 PWCNet 提取场景光流特征图，并与原视频帧融合作为基本单元，由多个基本单元组成连续基本单元作为模型的输入；利用 3DConv 和 ConvLSTM 模块进行时空特征的自主提取，3DSEblock 模块进行重要特征的筛选；最终，通过输入数据和自编码器重建视频块之间的重建误差，来判断视频是否出现异常行为。通过在 UCSD、Avenue 等公开数据集上进行验证，实验结果的定性和定量分析证明了本方法具有较好的性能。

关键词：异常检测；ST-3DCAE；特征融合

Video anomalous behavior detection based on 3D convolutional auto-encoder

LIAN Jing, HU Xing, HUANG Yingping
(School of Optical Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Video anomalous behavior detection is one of the current hot problems in the field of computer vision. However, the difficulty of specific definition of anomalous behavior makes it difficult to apply supervised learning-based two-class classification methods in this field. In this paper, we propose an unsupervised video anomaly detection model, called 3D self-encoder model based on spatio-temporal feature fusion (ST-3DCAE). First, the optical flow feature map of the scene is extracted with PWCNet and fused with the original video frame as the basic unit, and multiple basic units form a continuous basic unit as the model input; then, the 3DConv module and ConvLSTM module are used for autonomous extraction of spatio-temporal features, and the 3DSEblock module is used for screening of important features; finally, the reconstruction of the self-encoder through the input data and reconstruction error between video blocks to determine whether the video shows abnormal behavior. The model proposed in this paper was validated on publicly available datasets such as UCSD and Avenue, and the qualitative and quantitative analysis of the experimental results proved the better performance of this method.

【Key words】 anomaly detection; ST-3DCAE; feature fusion

0 引 言

智能视频监控系统能够广泛地应用于生产及生活领域的安全防护[1]。随着监控设备价格的降低，以“天网监控系统”为代表的监控系统已全方位深入到社会每个角落。视频异常行为代表一类相对罕见的行为，一般与常见的行为相冲突，具有一定的危险性，及时准确地检测视频异常行为对于维护社会安全具有重要意义。视频异常检测旨在通过计算机视觉，结合机器学习等方法，在时间和空间上定位视频序列中有违常规行为。例如，车辆逆行、打架、人群逃散等行为。合适的视频异常行为检测方法能够自动学习场景中的正常行为模型，并自动判断出与正常模型偏差较大的异常行为，不仅大大减少了劳动成本，而且有着较好的时效性和准确率。

不同于其它基于监督学习的视频行为识别任务，由于异常行为具有定义模糊性[2]、稀有性[3]、场景依赖性[4]等特点，所以监督学习不能很好的适用于视频异常检测。因此，大多数视频异常检测都是通过无监督学习的方式来实现的。文献[5]中首次运用深度去噪自编码器重建时空立方体，通过全连接层进行输出作为习得的视频事件表示；文献[6]使用卷积深度自编码器，通过卷积代替全连接层，以便更好地提取视频特征；文献[7]使用赢家通吃自编码器提取视频特征，并使用一类支持向量机进行分类等等。以上文献均使用了深度自编码器，并取得了较好的效果。但由于视频具有时间上的连续性，而上述方法仅考虑了空间上的特征提取，忽略了

基金项目：国家重点研发计划（2019YFB1705702）。

作者简介：连 靖（1994-），男，硕士研究生，主要研究方向：基于计算机视觉的异常行为检测。
通讯作者：胡 兴 Email：huxing@usst.edu.cn

收稿日期：2021-03-13

时间序列上的关联性。

对于视频预测任务，空间和时间特征的结合是不可或缺的。为了解决上述问题，本文提出了一种新的基于深度学习的方法。其主要思想是，通过实现正常场景下的空间多尺度特征与时间信息的融合来优化异常检测。该方法结构如图 1 所示。首先，利用 PWC-Net[8] 网络提取相邻帧的光流特征图，将 RGB 图与光流图进行加权相加形成合成特征图，作为编码模块的输入；然后，利用三维尺度上的 attention 模块，对经 3D 卷积层的特征图进行主要信息的增强和次要信息的抑制；利用 ConvLSTM 模块在时间建模方面的优势来记忆时态特征，完成时空信息融合；最后根据解码模块获得重建视频数据，并根据重建数据与输入数据之间的差异获得每个输入的正常分数。

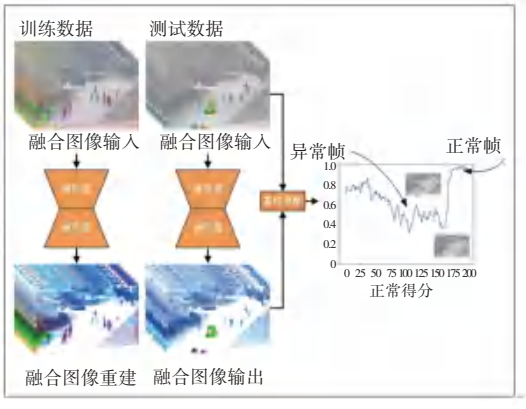


图 1　异常行为检测结构框图

Fig. 1　Block diagram of abnormal behavior detection

1　模型设计

本文所提出的框架如图 2 所示，该框架整体为编码-解码结构，主要由 3D 卷积模块、3DSEblock 模块和 ConvLSTM 模块组成。

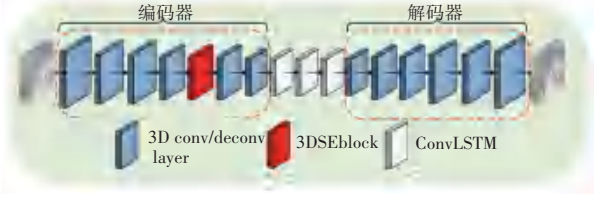


图 2　异常行为检测网络框架图

Fig. 2　Anomalous behavior detection network framework diagram

1.1　3D 卷积模块

2D 卷积使用空间卷积核对图像进行水平维度和垂直维度的特征提取，如公式(1)所示，只能提取图像的空间特征。而视频数据既有空间维度上的信息，也有时间维度上的信息，所以本文使用 3D 卷积核，同时对水平维度、垂直维度和时间维度进行卷积操作，不仅提取了视频的空间特征，还建立了空间特征在时间维度上的联系，如公式(2)所示。

$$a_{ij} = \sum_{m=1}^{K}\sum_{n=1}^{K} W_{mn} X_{i+m,\,j+n}, \tag{1}$$

$$a_{ijz} = \sum_{m=1}^{K}\sum_{n=1}^{K}\sum_{p=1}^{P} W_{mnp} X_{i+m,\,j+n,\,z+p}, \tag{2}$$

1.2　3DSEblock 模块

通常情况下，同一地点的视频数据会存在大量冗余信息，而过多的冗余信息会影响模型的学习。所以，本文针对如何对主要信息进行增强，对次要信息进行抑制进行了模型上的改进。受文献[9]中提出的 SEnet 注意力机制模块启发，本文引入 3DSEblock 模块，对视数据进行主要信息增强与次要信息抑制。

该模块主要由 Squeeze 和 Excitation 两部分组成。前者主要利用全局平均池化层统计每个通道的分布，来进行信息嵌入，如公式(3)所示；后者则通过学习得到一组权重，用于表示特征通道的重要程度，如公式(4)所示。

$$z_c = F_{sq}(u_c) = \frac{1}{T\times H\times W}\sum_{i=1}^{T}\sum_{j=1}^{H}\sum_{k=1}^{W} u_c(i,j,k), \tag{3}$$

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)), \tag{4}$$

$$y_c = F_{\text{scale}}(u_c, s_c) = s_c \cdot u_c. \tag{5}$$

其中，z_c 为第 c 个特征通道经过全局平均池化层所得到的输出值；T、H、W 分别表示该特征的帧数、高和宽；$u_c(i,\ j,\ k)$ 为第 c 个特征通道在位置 (i,j,k) 处的值；s 表示通过 Excitation 操作所得到的一组权重系数。由图 3 可知，W_1 和 W_2 分别为第一层和第二层全连接层的权重，此 2 层通过缩放因子 r 来控制计算的复杂程度；δ 为 ReLU 激活函数，用于提高模块的非线性程度；σ 为 sigmoid 激活函数，用于将权重正则到(0,1)。最后将该权重与对应输入相乘，得到最终的输出，具体实现如公式(5)所示。y_c 表示第 c 个特征通道 x_c 经过 3DSEblock 模块所得到的输出值；u_c 是 x_c 经 3D 卷积模块后所得到的结果；s_c 是通过 Excitation 操作所得到的一组权重系数，其系数越大，代表该特征通道越重要。

1.3　ConvLSTM 模块

ConvLSTM(The convolution Long Short-Term Memory)是在 LSTM(Long Short-Term Memory)的基础上改进而来。由于视频数据需要同时考虑时间维度上的连续性和空间维度上的关联性，而 LSTM 虽然能提取数据的时序信息，但会破坏数据的空间结构。为了能够同时提取时空特征，Shi[10] 提出了卷积层和

LSTM 单元相结合的 ConvLSTM 模块并用于降水预测中。相较于 LSTM,全连接层被卷积操作所替代,且有着局部连接、参数共享的优点,既有效地降低了参数量,又能够提取视频帧的局部特征。所以 ConvLSTM 模块更适合于视频流数据。ConvLSTM 结构如图 4 所示。其中,t 表示 t 时刻;X_t、H_t、C_t 分别表示 t 时刻的输入、隐藏状态及记忆单元;3 * 3 矩阵区域表示卷积核。

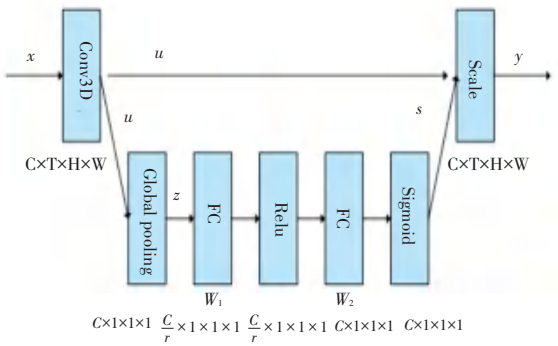


图 3 3DSEblock 模块结构图

Fig. 3 3DSEblock module structure diagram

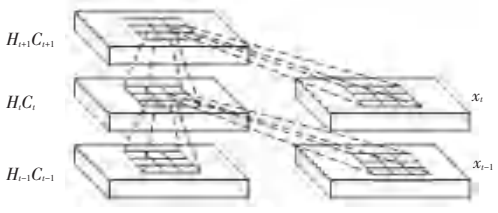


图 4 ConvLSTM 结构图

Fig. 4 ConvLSTM structure diagram

2 实验结果及分析

2.1 实验数据集

2.1.1 UCSD 数据集

UCSD 数据集[11]是视频异常检测最常用的数据集之一,记录了加州大学圣地亚哥分校两个不同人行道上的监控视频数据。其中,正常行为指行人在人行道上的行走,异常行为包括不规范的机动车辆行驶、单车行驶及滑板穿梭等行为。Ped1 数据集包含 34 个训练样本和 36 个测试样本,每个样本由 200 帧组成,每帧分辨率为 158X238 像素。Ped2 数据集包含 16 个训练样本和 12 个测试样本,每个样本由 120-200 帧组成,总帧数为 4 560 帧,每帧分辨率为 240X360 像素。

2.1.2 Avenue 数据集

Avenue 数据集[12]采集于香港中文大学某走廊的监控视频数据。其中,正常行为指行人在平行于地铁口的方向上行走,异常行为包括但不限于行人在非平行于地铁口的方向行走、投掷物品、奔跑、推自行车等行为。该数据集包含 16 个训练样本和 21 个测试样本,每个样本帧数各不相同,总帧数为 30 337帧,每帧分辨率为 360X640 像素。

2.2 评价标准

对异常检测任务而言,其最常见的评价层次为帧层次比较与像素层次比较[11]。前者只须判断帧内是否有异常,后者需要对异常进行空间定位。当检测出的异常区域与真实的异常区域重合率大于 40%时,该帧被认为是异常帧。为了在量上进行精确的比较,实验中经常用受试者工作特性曲线(ROC)下的面积(AUC),以及错误率(EER)等,来评估视频异常检测的性能。ROC 曲线由不同阈值下的真阳性率(TPR)和假阳性率(FPR)绘制,其中:

$$TPR = \frac{TP}{TP + FN}, \tag{6}$$

$$FPR = \frac{FP}{FP + TN}. \tag{7}$$

TP 和 FP 分别代表正确检测与错误检测的阳性样本,TN 和 FN 分别代表正确检测与错误检测的阴性样本。帧级标准由等差错率(EER)概括。其是在 ROC 曲线上 $FPR = TPR$ 时的值,而像素标准则通过检测率来概括。

此外,为了判定该模型是否有效,需要用测试集评估该模型是否能够捕捉到异常行为,所以本论文采用正常得分来衡量模型效果。理论上正常行为得分较高,异常行为得分较低。正常得分的计算方法如公式(8)、式(9)所示:

$$d(t) = \| x(t) - f_W(x(t)) \|_2, \tag{8}$$

$$s(t) = 1 - \frac{d(t) - d(t)_{\min}}{d(t)_{\max}}. \tag{9}$$

其中,$d(t)$ 表示第 t 帧的重构误差,即一帧图像中所有像素的重构误差;$x(t)$ 表示视频的第 t 帧;f_W 表示已训练好的模型;W 表示已训练模型的权重;$f_W(x(t))$ 表示视频第 t 帧在输入模型后得到的重构帧。输入帧和重构帧的二范数即为重构误差。公式(9)对重构误差进行了归一化。其中,$s(t)$ 表示第 t 帧的重构误差,$d(t)_{\max}$ 和 $d(t)_{\min}$ 分别表示输入帧中最大的重构误差和最小的重构误差。

2.3 实验环境及训练设置

实验的服务器配置包括:Intel Xeon E5-2667 v4 CPU、NVIDIA GeForce GTX 2080 和 11 GB 内存。实验之前,需要对原视频帧进行尺寸调整,数据中心化

以及数据增强。本文统一将视频帧大小调整为224x224，并进行数据中心化。即将训练集中每一帧的每个像素先转化到(0,1)，然后减去μ再除以σ。本文取 μ = [0.485,0.456,0.406]，σ = [0.229,0.224,0.225]。为提高模型的泛化能力，避免过拟合，将中心化之后的视频帧进行采样，步长分别为1、2、3，然后以每 8 帧为一视频块输入到模型中。

2.4　UCSD 数据集实验结果

表 1 和表 2 分别列出了 UCSD Ped1 数据集和 Ped2 数据集下采用 ST-3DCAE 模型与其它主流方法在 *EER* 和 *AUC* 这两个指标上的详细定量比较。其中，若 *EER* 越小，*AUC* 越大，说明检测方法越可靠。从表 1 和表 2 可以明显看出，本文提出的方法除逊于 AMDN，与 Conv-AE 性能相近之外，均优于其它传统算法。因为这两种方法也同样使用了以卷积自编码器为基础的模型结构，在提取特征方面具有一定的优势。

表 1　不同方法在 UCSDPed1 数据集上的性能比较

Tab. 1　Performance comparison of different methods on UCSDPed1 dataset

Approach	*EER* /%	*AUC*/%
Adam[13]	38.0	77.1
SF[14]	31.0	67.5
MPPCA[11]	40.0	66.8
AMDN[6]	16.0	92.1
Conv-AE[15]	27.9	81.0
ST-3DCAE	25.1	80.7

表 2　不同方法在 UCSDPed2 数据集上的性能比较

Tab. 2　Performance comparison of different methods on UCSDPed2 dataset

Approach	*EER* /%	*AUC* /%
Adam[13]	42.0	/
SF[14]	42.0	55.6
MPPCA[11]	30.0	69.3
AMDN[6]	17.0	90.8
Conv-AE[15]	21.7	90.0
ST-3DCAE	21.8	85.3

图 5 分别展示了上述两个数据集的 ROC 曲线图。为进一步证明该方法的可靠性，图 6 可视化了 UCSDPed1 数据集的正常得分，图 7 可视化了 USDPed2 数据集的正常得分。其中横轴为视频帧数，纵轴为正常得分，红色区域为 ground truth，意味着该区域发生了异常行为。理论上正常行为帧拥有较高的正常得分，异常行为帧具有较低的正常得分。

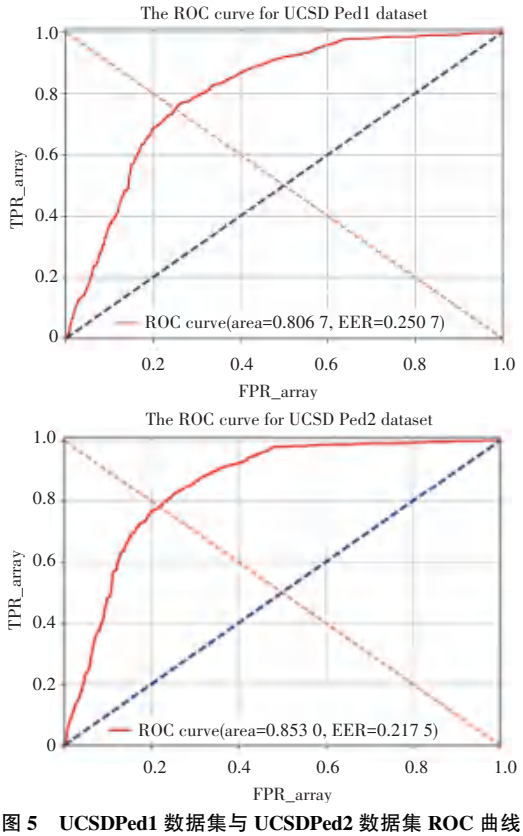
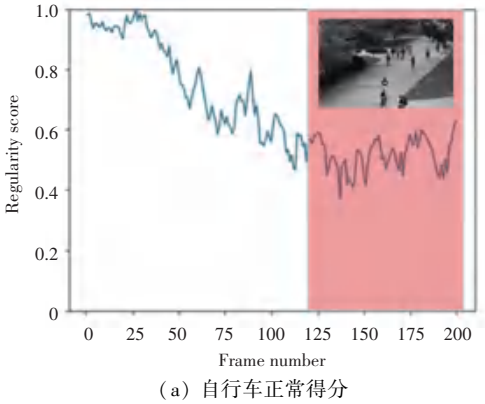


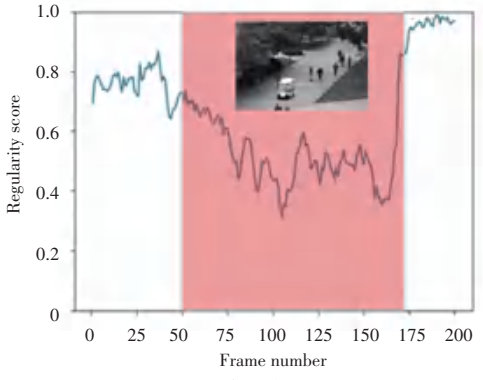
图 5　UCSDPed1 数据集与 UCSDPed2 数据集 ROC 曲线

Fig. 5　ROC curves for UCSDPed1 dataset and UCSDPed2 dataset



（a）自行车正常得分

（a）Normal score for bicycle

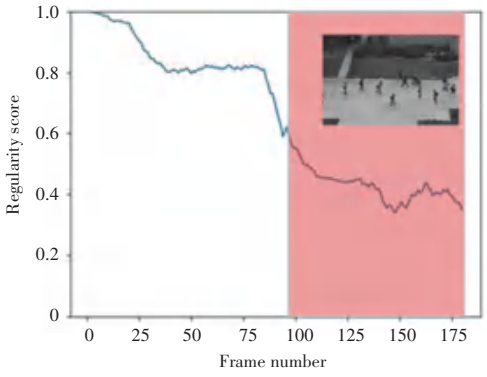


（b）汽车正常得分

（b）Normal score for car

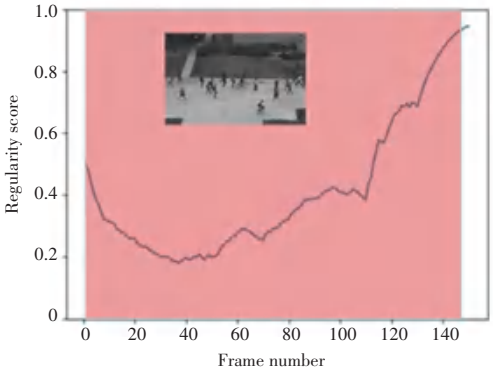
图 6　UCSDPed1 数据集下异常目标的正常得分

Fig. 6　Normal score plot for abnormal targets under UCSDPed1 dataset



(a) 单人骑车正常得分

(a) Normal score for single rider



(b) 多人骑车正常得分

(b) Normal score for muti-rider

图 7 UCSDPed2 数据集下异常目标(自行车)的正常得分

Fig. 7 Normal score plot for abnormal target (bicycle) under UCSDPed2 dataset

2.5 Avenue 数据集实验结果

表 3 列出了 Avenue 数据集下采用 ST-3DCAE 模型与其它主流方法在 *EER* 和 *AUC* 指标上的详细定量比较。从表 3 中可以明显看出本文提出的方法优于 Conv-AE,且与其它传统算法的性能相近,从而验证了 ST-3DCAE 能够较为有效地定位异常帧。图 8 展示了 Avenue 数据集的 ROC 曲线。为进一步证明该方法的可靠性,图 9 可视化了 Avenue 数据集的正常得分。

表 3 不同方法在 Avenue 数据集上的性能比较

Tab. 3 Performance comparison of different methods on Avenue dataset

Approach	*EER*/ %	*AUC*/ %
Conv-AE[6]	25.1	70.2
ConvLSTM-AE[16]	20.7	80.3
GMFC-VAE[17]	22.7	83.4
ConvAD-AE	22.3	84.2
ST-3DCAE	24.9	81.0

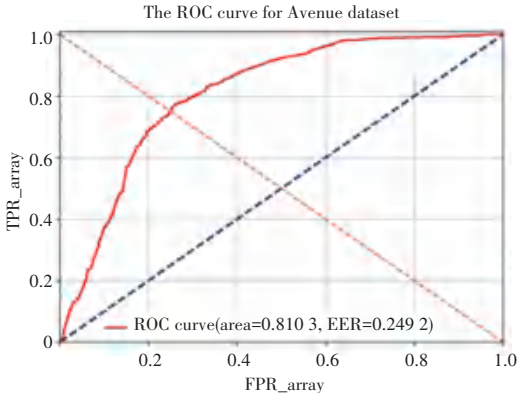
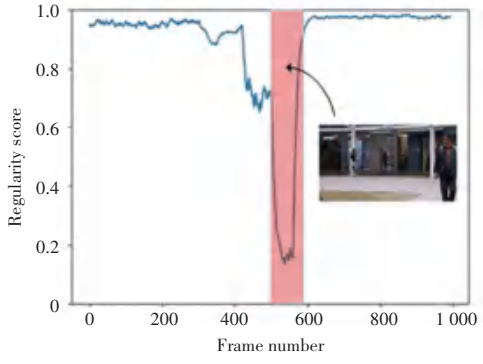


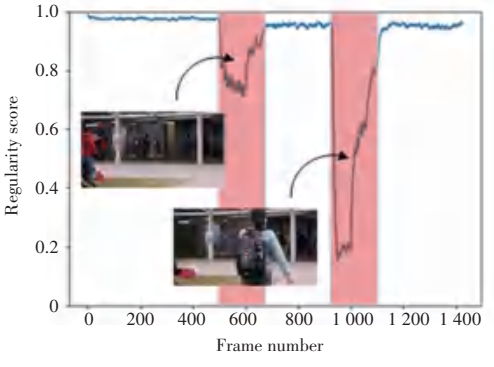
图 8 Avenue 数据集的 ROC 曲线

Fig. 8 ROC curves for the Avenue dataset



(a) 面向监控行走状态下正常得分

(a) Normal score in the state of facing the monitor walking



(b) 徘徊/背向监控行走状态下正常得分

(b) Normal score in the state of wandering/backing the monitor walking

图 9 Avenue 数据集下异常目标(行人)的正常得分图

Fig. 9 Normal score plot for abnormal targets (pedestrians) under Avenue dataset

3 结束语

对于目前大多数的异常检测任务使用的均为无监督学习模型,即使用正常样本来构建生成模型。然而,传统的无监督学习模型往往会造成假阳性率偏高。本文针对连续视频帧的特性提出了一种改进后的无监督学习模型。该模型使用了 3D 卷积模块和 ConvLSTM 模块充分提取了连续视频帧的特征,

并且使用了 3DSEblock 模块用于主要信息增强和次要信息抑制。实验结果表明该模型在异常行为检测这一任务中有着较为优秀的性能。当然，该模型也有着一定的局限性，即同一个场景仅对应一个固定参数的模型，若更换场景，模型必须重新训练。进一步将计划攻克这些难题，构建出一种在绝大多数场景下通用的异常行为检测模型。

参考文献

[1] PAUL M，HAQUE S M E，CHAKRABORTY S. Human detection in surveillance videos and its applications-a review[J]. EURASIP Journal on Advances in Signal Processing，2013(1)：1-16.

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

[3] 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，2018：6536-6545.

[4] CHONG Y S，TAY Y H. Modeling representation of videos for anomaly detection using deep learning：A review[J]. arXiv preprint arXiv：1505.00523，2015.

[5] XU D，YAN Y，RICCI E，et al. Detecting anomalous events in videos by learning deep representations of appearance and motion[J]，2017，156：117-127.

[6] 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，2016：733-742.

[7] TRAN H T，HOGG D. Anomaly detection using a convolutional winner-take-all autoencoder[C]// Proceedings of the British Machine Vision Conference 2017，2017.

[8] SUN D，YANG X，LIU M Y，et al. Pwc-net：Cnns for optical flow using pyramid，warping，and cost volume[C]//Proceedings of the IEEE conference on computer vision and pattern recognition，2018：8934-8943.

[9] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition，2018：7132-7141.

[10] XINGJIAN S，CHEN Z，WANG H，et al. Convolutional LSTM network：A machine learning approach for precipitation nowcasting[C]//Advances in neural information processing systems，2015：802-810.

[11] MAHADEVAN V，LI W，BHALODIA V，et al. Anomaly detection in crowded scenes[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition，2010：1975-1981.

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

[13] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv preprint arXiv：1412.6980，2014.

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

[15] XU D，RICCI E，YAN Y，et al. Learning deep representations of appearance and motion for anomalous event detection[J]，Computer Vision and Image Understanding，2015.

[16] CHONG Y S，TAY Y H. Abnormal event detection in videos using spatiotemporal autoencoder[C]// International Symposium on Neural Networks，2017：189-196.

[17] FAN Y，WEN G，LI D，et al. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder[J]. Computer Vision and Image Understanding，2020，195：102920.

(上接第 69 页)

重构活动等问题。本文算法还存在得到的 Hunk 不准确和多个最长公共子序列的部分问题，后续工作需进一步解决这方面问题，继续提高差异分析准确率，推广其到其他语言开源项目中的差异分析。

参考文献

[1] BAUM T，SCHNEIDER K，BACCHELLI A. On the Optimal Order of Reading Source Code Changes for Review[C]// 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE，2017. 329-340.

[2] TAO Y，DANG Y，TAO X，et al. How do software engineers understand code changes? - An exploratory study in industry[C]// Acm Sigsoft International Symposium on the Foundations of Software Engineering. ACM，2012.

[3] MYERS E W. AnO (ND) difference algorithm and its variations[J]. Algorithmica，1986，1(1-4)：251-266.

[4] JOKINEN P，UKKONEN E. Two algorithms for approxmate string matching in static texts[C]//International Symposium on Mathematical Foundations of Computer Science. Springer，Berlin，Heidelberg，1991：240-248.

[5] JOKINEN P，UKKONEN E. Two algorithms for approxmate string matching in static texts[C]//International Symposium on Mathematical Foundations of Computer Science. Springer，Berlin，Heidelberg，1991：240-248.

[6] 石伟，杨春花. 基于变更块的代码重构模式展示—以抽取方法为例[J]. 智能计算机与应用，2019，9(3)：85-88.

[7] BRITO R，VALENTE M T. RAID：Tool Support for Refactoring-Aware Code Reviews[J]. arXiv preprint arXiv：2103.11453，2021.

[8] FLURI B，WURSCH M，PINZGER M，et al. Change distilling：Tree differencing for fine-grained source code change extraction[J]. IEEE Transactions on software engineering，2007，33(11)：725-743.

[9] CHAWATHE S S，RAJARAMAN A，GARCIA-MOLINA H，et al. Change detection in hierarchically structured information[J]. Acm Sigmod Record，1996，25(2)：493-504.

[10] JACCARD P. The distribution of the flora in the alpine zone[J]. New Phytologist，2010，11(2)：37-50.

[11] YANG C，WHITEHEAD E J. Pruning the AST with Hunks to Speed up Tree Differencing[C]// 2019 IEEE 26th International Conference on Software Analysis，Evolution and Reengineering (SANER)，Hangzhou，China，2019：15-25.