引文格式:潘文康,邵振峰,廖明,等.利用深度时空自编码网络与多示例学习进行船只异常事件检测[J].武汉大学学报(信息科学版),2024,49(7):1109-1119.DOI:10.13203/j.whugis20220121

Citation:PAN Wenkang,SHAO Zhenfeng,LIAO Ming,et al.Ship Abnormal Event Detection with Deep Spatiotemporal Autoencoder Network and Multi-instance Learning[J].Geomatics and Information Science of Wuhan University,2024,49(7):1109-1119.DOI:10.13203/j.whugis20220121

利用深度时空自编码网络与多示例学习进行船只异常事件检测

潘文康[1] 邵振峰[1] 廖 明[2] 李先怡[3] 宋 杨[4]

1 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

2 江西省自然资源事业发展中心,江西 南昌,330025

3 珠海欧比特宇航科技股份有限公司,广东 珠海,519080

4 广州市城市规划勘测设计研究院,广东 广州,440100

摘 要:异常事件检测是交通安全防控的重要支撑技术,也一直是信息科学领域的研究热点。提出了利用深度时空自编码网络与多示例学习进行船只异常事件检测的方法,针对目前无法为模型训练提供精确帧级别标注的问题,引入多示例学习模型,将视频作为包,并将视频片段作为包中的示例,通过网络自动学习一个深度异常排序模型,该模型能预测异常视频片段的分数。同时,在特征提取方面,提出了深度时空自编码网络,在空间自编码器中,为了获取更精确的红绿蓝特征,将解码器中的上采样层替换为像素重组层。在时间自编码器中,为了突出运动变化较大的区域,引入基于方差的注意力机制,使快速移动的物体有更大的运动损失,有利于检测出异常事件。还构建了一个新的大规模的船只视频数据集,包括100个真实场景的监控视频以及5类真实的异常事件,分别为海面逗留、非港口靠岸、非港口离岸、超速和越界。该数据集可用于模型的训练与测试。实验结果表明,相比传统的双流网络以及基于图像重构的检测方法,所提出的基于深度时空自编码网络与多示例学习的方法的异常事件检测精度由71.7%提升为82.4%,表明了其在船只异常事件检测上的有效性。

关键词:船只异常事件检测;深度时空自编码网络;多示例学习;船只视频数据集

中图分类号:P237 **文献标识码**:A **收稿日期**:2022-08-10

DOI:10.13203/j.whugis20220121 **文章编号**:1671-8860(2024)07-1109-11

Ship Abnormal Event Detection with Deep Spatiotemporal Autoencoder Network and Multi-instance Learning

PAN Wenkang[1] *SHAO Zhenfeng*[1] *LIAO Ming*[2] *LI Xianyi*[3] *SONG Yang*[4]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

2 Natural Resources Development Center of Jiangxi Province, Nanchang 330025, China

3 Zhuhai Orbita Aerospace Science & Technology Co. Ltd, Zhuhai 519080, China

4 Guangzhou Urban Planning and Design Survey Research Institute, Guangzhou 440100, China

Abstract: **Objectives:** RGB(red green blue) and motion features are very important for ship video abnormal event detection. We need to extract these features in video more accurately and apply them to the detection of abnormal events in ship video. Meanwhile, due to the huge cost of frame-level annotations, we also need to solve the problem of not providing frame-level annotations in the model training stage, but using video-level annotations for model training. In addition, we also need to solve the problem of the scarcity of ship video abnormal event database. **Methods:** We acquire a large number of surveillance videos of ships on the sea surface and construct a data set of abnormal events of ship video after processing. Also, we propose

基金项目:湖北省重点研发计划(2022BAA048);山西省科技重大专项计划(202201150401020)。

第一作者:潘文康,硕士,主要研究方向为视频异常事件检测。panwenkang@whu.edu.cn

通讯作者:邵振峰,博士,教授。shaozhenfeng@whu.edu.cn

a ship abnormal event detection model based on deep spatiotemporal autoencoder network and multi-instance learning, using a deep multi-instance ranking framework, without obtaining frame-level annotations, only video-level information is needed. In addition, in the spatial autoencoder, in order to obtain more accurate RGB features, the deconvolution layer in the decoder is replaced by a pixel shuffle layer. In the temporal autoencoder, in order to highlight the regions with large motion variation, a variance-based attention mechanism is introduced, so that fast-moving objects have a larger motion loss. **Results:** We compare the proposed method with the two benchmark methods and a previous state-of-the-art method. The experimental results show that the proposed method has higher detection accuracy. In addition, we observe that variance-based attention can significantly improve the detection effect of fast motion, such as unexpected stopping and overspeed. **Conclusions:** This shows that RGB and motion features play an important role in ship abnormal event detection and also proves the necessity of multi-instance ranking model.

Key words: ship abnormal event detection; deep spatiotemporal autoencoder network; multi-instance learning; ship video dataset

态势感知是一种动态、整体地洞悉安全风险的能力,是以安全大数据为基础,从全局视角提升对安全威胁的发现识别、理解分析、响应处置能力的一种方式,最终是为了决策与行动,是安全能力的落地[1-2]。近年来,随着沿海地区面临的海上安全问题日益增多,海域态势感知也逐渐成为研究的热点。海域态势感知被国际海事组织定义为“对可能影响安全、经济、或环境的与海事领域相关的任何事物的有效理解”[3]。

船只异常事件检测作为海域态势感知的关键技术之一[4-6],也一直是信息科学领域研究的热点。其主要目的是从正在行驶的船只视频中提取出船只在行驶过程中的普遍行为特征,从而能识别检测出与总体运动特征差异较大的船只个体。通常来说,船只的异常行为包括海面逗留、非港口离岸、非港口靠岸、超速等。当船只在行驶过程中出现这些异常行为时,往往预示着危险,如偷渡、两船相撞、非法捕鱼等,而且船只的运动轨迹会发生明显的变化,及时准确地从船只的运动特征中检测出异常行为,有利于帮助监管人员识别危险,从而能采取相应的措施应对危险。

现实世界中的异常事件是复杂多样的,无法列举出所有可能的异常事件,也无法对视频进行精确的帧级别标注。因此,根据文献[7],目前主流的异常检测方法都是通过对正常数据进行特征重构来解决数据标注有限的问题。这些方法通过学习一个自动编码器或U-Net[8]来重构正常事件或预测真实帧,从而检测出视频中是否有异常。基于重构的异常检测方法[9-11]将手工特征或者视频帧作为模型的输入,提取高级的特征表示用来对正常事件建模,并通过最小化重构误差来学习正常事件中的时间规律。由于这些模型只学习正常事件中的模式,异常模式会导致更大的重构误差,因此,可以通过重构误差来检测异常事件。文献[12]通过二维卷积自编码器来对特征进行降维并学习时间规律性。文献[13]通过使用相邻帧的时间相干性来训练自动编码器网络。文献[14]通过引入无标签监督,使用约束学习,并且结合物理和领域知识,共同解决目标跟踪的计算机视觉任务。但是,由于深度神经网络的高泛化能力,异常事件的重构误差不一定会比正常事件的重构误差大,因此文献[7]提出了一种基于预测的异常事件检测方法,该方法使用U-Net架构从之前的连续帧预测未来的下一帧,然后将预测结果与真值进行比较来判断是否为异常事件。但是该方法没有充分利用视频中的时间信息。

为了充分利用视频中的空间和时间信息,文献[15]首次提出双流网络,即红绿蓝(red green blue, RGB)流和光流,并且通过将两个流提取的特征进行融合来对视频中的动作进行分类。由于异常事件可以通过外观或运动来检测,因此文献[16]提出一个具有两个网络分支的时空模块来检测异常事件,该方法将目标的外观和运动模式联合建模,展示了双流架构在复杂场景中建模的有效性。文献[17]引入用于视频异常事件检测的双流架构,将光流表示的图像块和运动特征作为两个独立的网络输入,分别提取外观特征和运动特征,将这两个特征融合并计算相应的异常分数。文献[18]利用两个生成器网络来学习人群行为的正常模式,其中一个生成器网络获取输入帧来生成光流图像,另一个生成器网络从光流中重建帧。虽然上述方法都很好地利用了视频

中的空间和时间信息,但是这些方法无法准确区分出视频中的普通目标和快速移动的目标,不利于检测出异常。

总的来说,上述方法无论是在检测精度上还是在特征提取上都存在不足。因此,为了解决模型过度依赖帧级别标注以及无法精确提取时间和空间特征的问题,本文提出了一种基于深度时空自编码网络和多示例学习的异常事件检测方法,利用深度多示例排序框架解决了无法获取帧级别标注的问题,只需要获取视频级别的标注。另外,本文结合自编码网络以及双流网络的优点,提出了深度时空自编码网络,该网络能更好地提取视频的时空特征。在空间自编码器中,为了使重构的图像与原始图像更相似,获取更精确的RGB特征,将解码器中的上采样层替换为像素重组层。在时间自编码器中,引入基于方差的注意力机制,突出光流图中运动变化较大的区域,赋予快速运动目标更大的运动损失。实验结果表明,深度时空自编码网络提取的时空特征能较大提升模型对异常事件检测的能力。

1　深度时空自编码网络

C3D[19]和I3D[20]等基于卷积的特征是使用3D卷积在多个视频帧上计算得到的,使用这种方法得到的特征包含了时间序列信息,该方法的性能要优于其他仅仅使用RGB特征的方法。有研究表明[21]仅仅使用卷积特征的模型无法达到最优的性能,而将卷积特征和光流特征相融合,往往能达到最优的检测效果。双流网络[15]是行为检测和识别最常用的一种网络。它由空间流网络和时间流网络构成,空间流网络能获取图像的RGB信息,时间流网络提取出光流信息,该网络在异常事件检测方面也有较好的性能。

基于上述研究,本文提出了深度时空自编码网络用于船只的异常事件检测,如图1所示。它由空间流和时间流组成,空间流由空间自编码器构成,输入为视频帧图片,图片经过空间自编码器能提取运动目标的RGB特征。时间流由时间自编码器构成,输入为运动目标的运动特征,比如光流。该方法融合空间流和时间流提取的运动目标特征,将该特征用于异常事件检测能提高检测的精度。在空间自编码器中,本文将原有的反卷积层用像素重组层替代,提高了图片帧经过重构后的空间分辨率,保留了输入图片更多的信息。在时间自编码器中,在编码阶段引入基于方差的注意力模块,有助于检测出快速移动的目标。

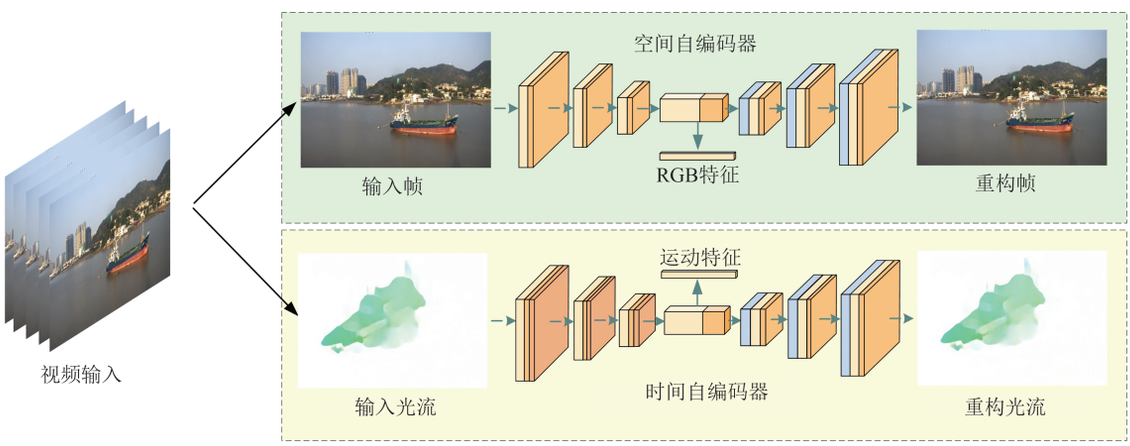

图 1　深度时空自编码网络

Fig. 1　Deep Spatiotemporal Autoencoder Network

1.1　空间自编码器

由于视频中大部分对象都具有关联性,因此,视频中的异常目标与某些特定对象也存在关联,比如在某些特定场景下才会发生某一类异常事件,所以这些特定对象的RGB特征同样十分重要。为了检测出具有场景和外观等空间特征的异常对象,本文提出了基于像素重组的空间自编码器,网络结构如图2所示。

空间自编码器的输入为视频的每一帧,分辨率大小为256×256像素,空间自编码器共分为7层,分别为3层编码器、1层瓶颈层、3层解码器。编码器部分是由2D卷积层和ReLU激活函数组成,并且步幅为2来降低特征分辨率。为了在解码后减少棋盘效应,输出高质量的图片,用于获取高质量的RGB特征,解码器将上采样层替换为像素重组层,解码器由3部分组成,分别为像素重

组层、2D卷积层和ReLU激活函数。瓶颈层中使用全局平均池化得到RGB特征。

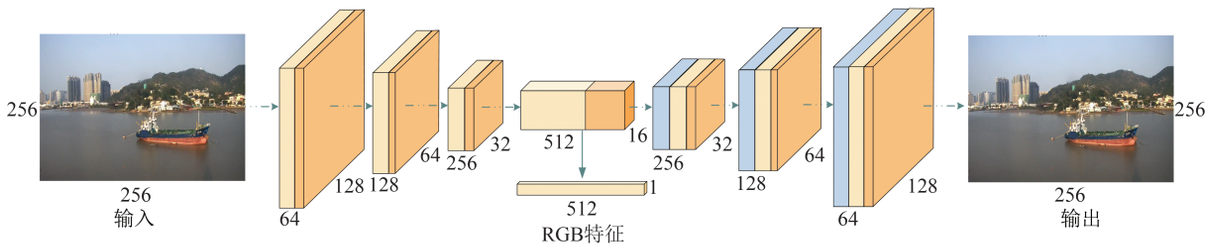

图2 空间自编码器

Fig. 2 Spatial Autoencoder Network

首先,将视频的某一帧x_i输入到空间自编码器网络中。然后,经过编码器能提取到输入帧的RGB特征,记为a_f,a_f包含了用于帧重建的基本空间信息。最后,经过解码器将RGB特征重建得到输出帧,记为$\bar{x}_i$。计算公式为:

$$a_f = E_m\left(x_i;\theta_e^m\right) \tag{1}$$

$$\bar{x}_i = D_m\left(a_f;\theta_d^m\right) \tag{2}$$

式中,E_m表示编码器;D_m表示解码器;θ_e^m表示空间编码器的参数;θ_d^m表示空间解码器的参数。

在训练阶段,为了训练空间自编码器学习RGB特征空间中的规律,最小化输入帧x_i和输出帧$\bar{x}_i$之间的重构误差。使用l_1重构误差,空间自编码器的损失函数l_1定义为:

$$l_1 = \left| x_i - \bar{x}_i \right| \tag{3}$$

1.2 时间自编码器

运动特征作为视频中的重要特征,也是异常事件检测的关键。由于光流是使用最广泛的运动特征,所以将光流作为自编码器的输入。与C3D类似,首先将一个视频处理片段设置为16帧,并将这16帧的分辨率调整为256×256像素,然后使用PWC-Net[22]来计算相邻帧之间的光流。每个光流图有两个通道,一个表示水平方向的移动,另一个表示垂直方向的移动。因此,运动增强网络的最终输入是15个尺寸为2×256×256的光流图F。

考虑到网络模型的效率问题,本文将时间自编码器设计成7层,如图3所示,分别由3层编码器、1层瓶颈层、3层解码器组成。编码器部分由2D卷积层、ReLU激活函数以及方差注意力模块组成。解码器部分与空间自编码器的解码器类似,共由3部分组成,分别为像素重组层、2D卷积层和ReLU激活函数。瓶颈层中使用全局平均池化得到运动特征。该网络可以使用l_1重构误差以一种无监督的方式在目标数据集上进行训练:

$$l_1 = \left| F - \widetilde{F} \right| \tag{4}$$

式中,$\widetilde{F}$是重建的光流图。

对于每一个16帧的视频处理片段,执行前向传递直到瓶颈层,并进行全局池化操作以导出维度为512×1的特征,该特征即为用于异常检测的运动特征。

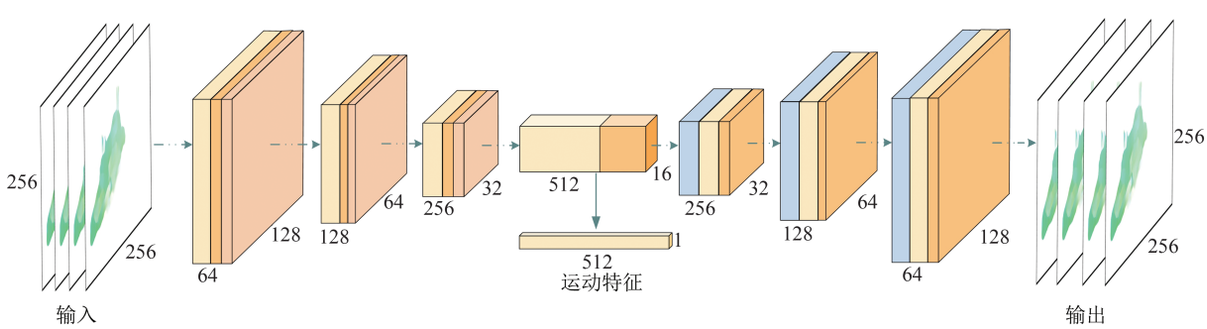

图3 时间自编码器

Fig. 3 Temporal Autoencoder Network

在视频中,如果出现异常事件,往往会有较大的运动变化。基于此特性,本文在时间自编码器中设计了一种基于方差的注意力,用来自动给视频中运动变化较大的部分分配更大的权重。

由于本文提出的时间自编码器的编码器由3个2D卷积块组成,因此特征图的每个位置都包含跨通道的局部运动信息,它类似于3D卷积,3D卷积包含了沿时间轴的运动信息,然而2D卷积在特征通道中也包含了这些信息。因此,对于运动变化较大的区域,嵌入层的方差会更高,所以

可以直接计算跨通道特征的平均值,然后计算每个位置的方差,如图 4 所示。

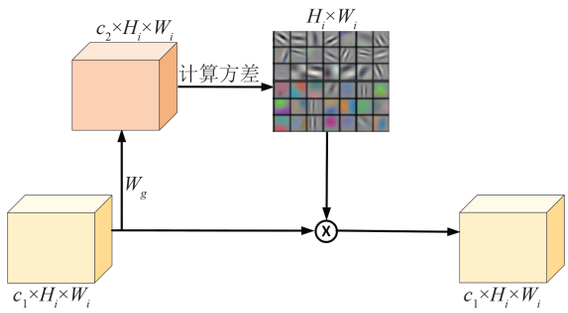

图 4　基于方差的注意力模块

Fig. 4　Variance Attention Module

考虑到时间编码器的第 i 块中嵌入的运动特征 z_m^i,基于方差的注意力模块首先需要将嵌入的运动特征输入到卷积层 f_n 中:

$$f_n(h,w)=W_g z_m^i(h,w) \tag{5}$$

式中,$h\in(0,H_i]$ 且 $w\in(0,W_i]$,H_i 和 W_i 分别为第 i 块的特征映射的行数和列数;W_g 表示卷积过滤器的权重参数,使用该卷积过滤器来得到输入特征的嵌入层。

先计算沿特征维的方差,然后沿空间维进行归一化运算,最后得到相应的权重图 g_n:

$$v(h,w)=\frac{1}{D}\sum_{d=1}^{D}\left(f_n(h,w,d)-\frac{1}{D}\sum_{d=1}^{D}f_n(h,w,d)\right)_2 \tag{6}$$

$$g_n(h,w)=\left[\frac{\exp\left(v(h,w)\right)}{\sum_{h=1,w=1}^{H,W}\exp\left(v(h,w)\right)}\right]_2 \tag{7}$$

式中,$d\in(0,D]$,D 为划分的块数;$v(h,w)$ 表示特征图在空间位置上的方差。

由于基于方差的注意力模块可以突出运动变化较大的区域,同时抑制不相关的静态区域,这些与快速移动高度相关的异常物体,即快速奔跑的行人,将会得到更大的运动损失。这有助于检测出快速移动的异常事件。

2　多示例排名模型

由于视频中异常事件的精确时间位置未知,因此不能像标准分类问题那样简单地学习异常模式,相反可以将其视为多示例学习问题。利用多示例学习模型,在船只异常事件检测时不需要获取精确的视频帧级别的标注,只需要视频级别的标注即可。

如图 5 所示,在多示例学习中,如果一个视频中包含异常事件,则该视频被标记为正样本,如果一个视频中没有异常事件,则该视频被标记为负样本。首先将正样本表示为正包,记为 B_a,然后将正样本中的视频按一定的时间间隔进行切分,得到的视频片段即为正包中的示例,记为 $(a^1,a^2,\cdots,a^m)$,m 为包中示例的个数,在正包中的示例中至少有一个包含异常事件。类似地,将负样本表示为负包 B_n,负包中的示例表示为 $(n^1,n^2,\cdots,n^m)$,在负包中,没有一个示例包含异常事件。在模型训练时,将每个视频分成 32 个片段,这就意味着每个包中包含 32 个示例。

根据之前的方法[23],将异常事件检测看作一个异常分数的回归问题。由于存在异常事件的视频片段比没有异常事件的视频片段具有更高的异常分数,因此定义的排名损失为:

$$f(V_a)>f(V_n) \tag{8}$$

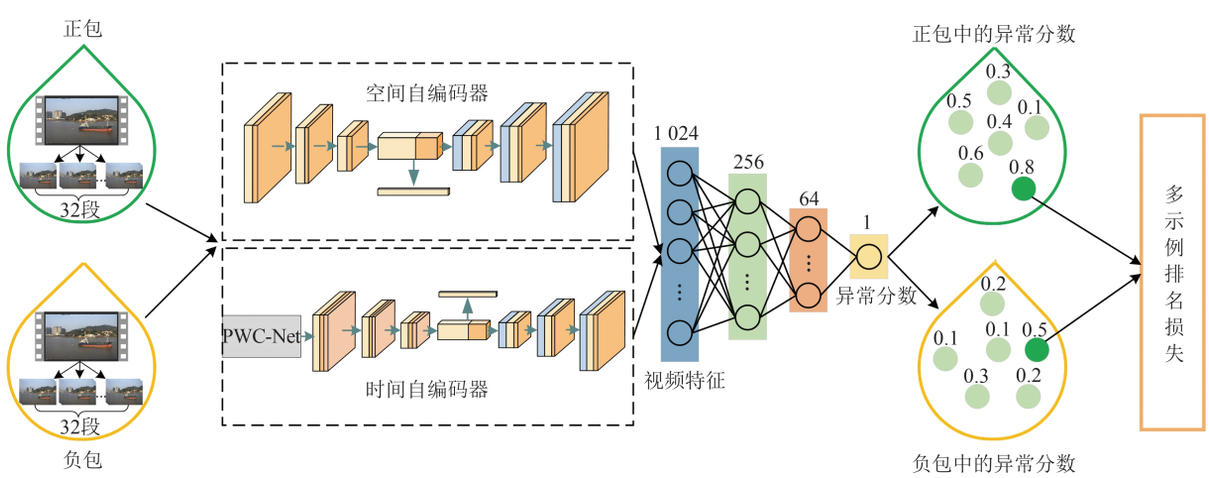

图 5　整体框架

Fig. 5　Overall Framework

式中,V_a和V_n分别表示异常和正常的视频片段;f是计算视频片段的异常分数的函数,取值范围为0到1,f由3个全连接层以及两个tanh激活函数组成,3个全连接层分别有256个单元、64个单元以及1个单元。该公式表明存在异常事件的视频片段的异常分数会更大。由于只有视频级别的标注,因此将上述排名损失加以改进,提出了多示例排名损失:

$$\max_{i\in B_a} f\left(V_a^i\right) > \max_{i\in B_n} f\left(V_n^i\right) \tag{9}$$

式中,V_a^i和V_n^i分别表示第i个异常和正常的视频片段;$f\left(V_a^i\right)$和$f\left(V_n^i\right)$分别表示第i个异常和正常的视频片段的异常分数;max指正包和负包中异常分数的最大值。这个排名损失的涵义是正包中异常分数的最高值比负包中异常分数的最高值要大。

3 船只异常事件检测实验

3.1 数据集

之前用于视频异常事件检测的数据集[7,23]都是基于行人或者车辆的,缺少基于船只的视频异常事件数据集。由于行人或者车辆的运动行为模式和船只的运动行为模式有很大不同,所以无法将基于行人或者车辆数据集训练出的模型直接用于船只的异常事件检测。因此,本文通过调研海关和边防的需求,从横琴海域获取了大量的船只视频数据,构建了基于船只的视频异常事件数据集(ship abnormal event dataset, SAED)。该数据集包含了100个真实场景的监控视频,视频时长为60 s,帧率为25帧/s,共有162 612帧。其中一半包含异常行为事件,另一半则只包含正常事件。该数据集包含5类真实的船只异常事件,分别为逗留、超速、非港口离岸、非港口靠岸、越界。在正常事件中,船只都是在海面上朝一个方向匀速行驶;在逗留事件中,船只在整个视频中静止不动;在超速事件中,船只相对其他行驶的船只速度明显更快;在非港口离岸和靠岸事件中,船只突然朝岸边靠近或远离;在越界事件中,船只突然变换航道。船只数据集中事件类别分布如表1和图6所示。本文中,该数据集将用于所有方法模型的训练和测试。

表1 船只数据集中事件类别分布

Tab. 1 Abnormal Event Category Distribution in Ship Dataset

类别	正常事件	海面逗留	超速	非港口离岸	非港口靠岸	越界
数量/个	53	14	8	7	8	10
帧总数	89 650	22 500	12 105	10 450	13 004	14 903
占比/%	53	14	8	7	8	10

3.2 评价指标

使用受试者工作特征(receiver operating characteristic, ROC)曲线和相应的曲线下面积(area under curve, AUC)以及等错误率(equal error rate, EER)作为评价指标[24],评估本文提出的检测方法在船只数据集上的性能。在ROC图像中,横轴为假阳性率,纵轴为真阳性率,ROC曲线可以直观地反映出分类性能的优劣[25]。AUC是ROC曲线与横轴所围区域的面积,是一种评价二分类型模型好坏的指标,其数值越大,模型性能越好。错误接受率和错误拒绝率相等时,它们的值即为EER。假阳性率F和真阳性率T的计算公式为:

$$F=\frac{R_{\mathrm{FP}}}{R_{\mathrm{FP}}+R_{\mathrm{TN}}} \tag{10}$$

(a) 逗留 (b) 超速 (c) 非港口离岸 (d) 非港口靠岸 (e) 越界

图6 船只异常事件类型

Fig. 6 Types of Ship Abnormal Event

$$T=\frac{R_{\mathrm{TP}}}{R_{\mathrm{TP}}+R_{\mathrm{FN}}} \tag{11}$$

式中,TP(true positive)表示真阳例,R_{TP}即正样本中被分类器预测为正样本的个数;FP(false positive)表示假阳例,R_{FP}即负样本中被分类器预测为正样本的个数;TN(true negative)表示真阴例,R_{TN}即负样本中被分类器预测为负样本的个数;FN(false negative)表示假阴例,R_{FN}即正样本中被分类器预测为负样本的个数。

3.3　实现细节

本文使用PyTorch[26]框架来训练提出的模型。首先,对于深度时空自编码网络,空间自编码器与时间自编码器训练过程类似。空间自编码器模型训练时,随机选取一帧,从选取帧中随机裁剪出一个256×256像素的子图像,然后进行水平翻转和色彩抖动。时间自编码器随机选择视频中连续的16帧作为一个处理片段,并将该处理片段作为PWC-Net的输入来计算光流。批量设置为32,优化器采用Adagrad,初始学习率设置为0.005。在训练阶段,网络的迭代次数设置为5万次,并且在迭代到2.5万次和4万次以及5万次时分别将学习率减半。其次,对于多示例排名模型,首先将每个视频分成32个不重叠的片段,在每个片段中,计算每个非重叠的16帧的特征,如果该段包含多个16帧处理片段,取所有特征的平均值,然后进行归一化。因此,对于每个视频,都能得到一个维度为32×512的运动特征,为了训练多示例排名模型,随机选择30个正包和30个负包作为批量,并且使用初始学习率为0.001的Adagrad优化器。同时,将模型的迭代次数设置为1万次,并在4 000次和8 000次时分别将学习率减半。

3.4　实验结果

如表2所示,将本文提出的方法和目前检测性能最好的基于图像重构的方法[27],以及两个基准方法[7,23]进行比较。文献[27]引入一个存储模块来记录正常事件的运动模式,并提出了特征的紧致性和分离性损失来训练存储模块,保证了存储模块分辨不同事件的能力。文献[7]提出了使用生成对抗模型预测图像的下一帧,并将预测帧与真值进行比较来检测异常事件。文献[23]引入了多示例学习模型,用于异常事件检测。由表2可知,所提方法与文献[27]方法相比,检测精度由71.7%提升为82.4%,并且与文献[23]方法和文献[7]方法相比,检测精度分别提升了23.2%和12.1%,表明了本文方法在异常事件检测上的有效性。另外,关于不可预料停船和超速这种有快速运动的一类,运动感知特征对这一类的检测效果的提升也很显著。这证明了运动增强网络学习到的运动感知特征的有效性。

表2　本文方法与其他方法在SAED上的实验结果对比/%

Tab. 2　Comparison of Experimental Results of the Proposed Method with Other Methods on SAED/%

方法	EER	AUC
文献[23]	22.3	59.2
文献[7]	20.7	70.3
文献[27]	18.5	71.7
本文方法	16.2	82.4

为了将实验结果可视化,将不同实验的ROC曲线绘制出来,如图7所示。由图7可以发现,本文所提方法明显优于其他几种方法,尤其是在假阳性率低于0.2的情况下,所提方法的真阳性率明显高于其他方法。从整条曲线来看,所提方法的真阳性率也是最高的。

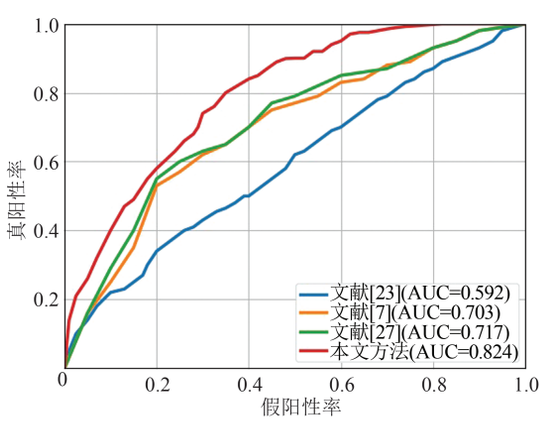

图7　不同异常检测方法的ROC

Fig. 7　ROC of Different Abnormal Detection Methods

在定性评价方面,图8展示了本文方法在SAED上的检测效果。蓝色线表示模型生成的当前帧的异常分数,淡红色区域表示异常事件出现的区间。可以看出,对于异常帧,本文方法能够通过生成高的异常分数来提供准确和及时的检测,对于没有发生异常的正常帧,生成的异常分数始终很低。这进一步表明了本文方法在船只异常事件检测方面的有效性。

3.5　消融实验

消融实验重在研究每一个模块对模型性能的影响,包括像素重组层、方差注意力机制以及结合空间信息和时间信息的方法。结合不同的模块在SAED数据集上进行实验。首先单独研究RGB特征和运动特征,再分别引入像素重组层

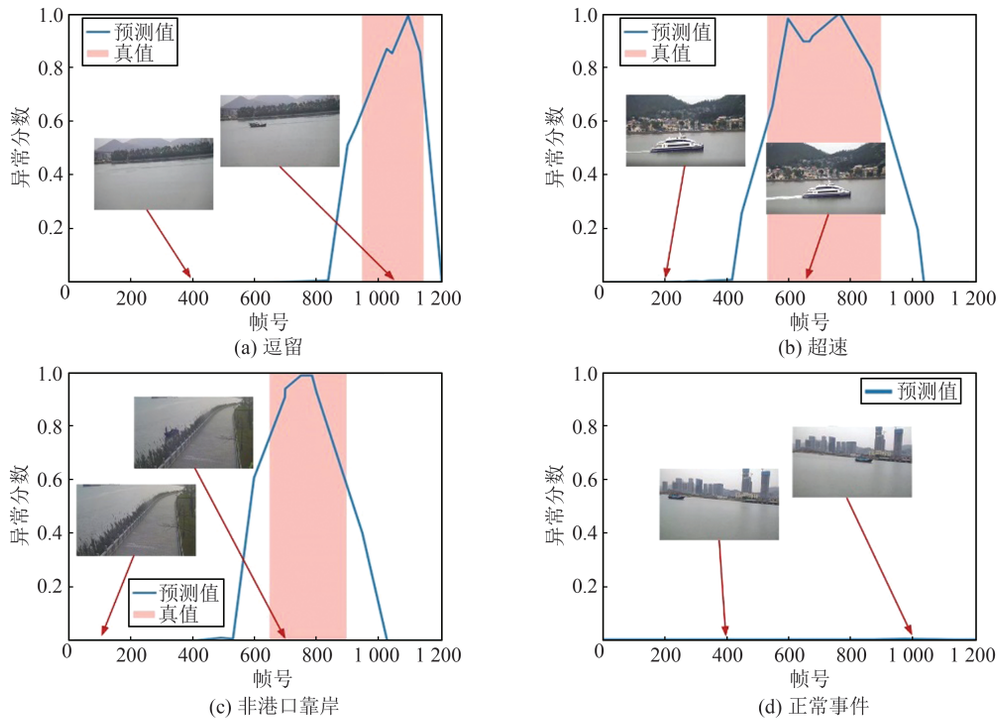

图8 船只异常事件检测可视化结果

Fig. 8 Visual Results of Ship Abnormal Event Detection

和方差注意力机制,然后将RGB特征和运动特征结合进行实验,最后一项实验包含所有模块。

表3表明了每个模块的有效性。由表3可知,与RGB特征相比,运动特征对于视频异常事件检测更为重要。当将RGB特征和运动特征相结合时,性能提升了0.9%。在此基础上,引入方差注意力机制,性能提升了1.9%。进一步引入像素重组层,性能提升了1.1%。

表3 各模块对模型性能的影响

Tab. 3 The Effect of Each Module on Model Performance

模块	实验1	实验2	实验3	实验4	实验5	实验6	实验7
RGB特征	√	√		√		√	√
像素重组层		√					√
运动特征			√	√	√	√	√
方差注意力机制					√	√	√
AUC/%	75.1	76.2	78.5	79.4	80.4	81.3	82.4

3.6 像素重组层对模型性能的影响分析

本节提出了基于像素重组的空间自编码器网络来提取视频中目标的RGB特征,将解码器中的反卷积层替换为像素重组层,大大提高了重构帧的分辨率,解决了插值和反卷积出现人工痕迹的问题,并且也消除了棋盘效应。由于RGB提取的质量直接影响重构帧的质量,因此为了评估利用改进后的空间自编码器对视频RGB特征的提取效果,分析对比了输入帧和重构帧在模型改进前后的差异。

通过样本集训练后的模型,从测试集中选取样本进行目标提取和分析。为了定量评价输入帧和重构帧之间的差异,使用峰值信噪比(peak signal-to-noise ratio, PSNR)来衡量输入帧和重构帧之间的相似性。PSNR是一种基于对应像素点之间误差的指标,是最常见、应用最广泛的图像客观评价指标。其计算公式为:

$$P(x_i,\bar{x}_i)=10\lg\frac{{x_{\max}}^2}{\frac{1}{N}\sum_{i=0}^{N}(x_i-\bar{x}_i)^2} \tag{12}$$

式中,P为PSNR;$x_{\max}$表示输入帧的最大像素值。PSNR值越大表示重构帧失真越小,重构得到的图片分辨率越高,获取到的RGB特征越好。

实验结果如图9所示。图9(a)为空间自编码器网络未改进时的重构图,即在解码阶段使用反卷积层得到的结果,将该模型记为DeAE;图9(b)

为使用ResNet块和反卷积层的网络得到的重构图片,将该模型记为ResAE;图9(c)为改进后的网络得到的输出结果,将该模型记为PSAE。图9中还展示了3种网络的输出与输入之间的PSNR值。由图9可知,改进后的空间自编码器在解码后得到的图像质量更高,具有更多的特征。

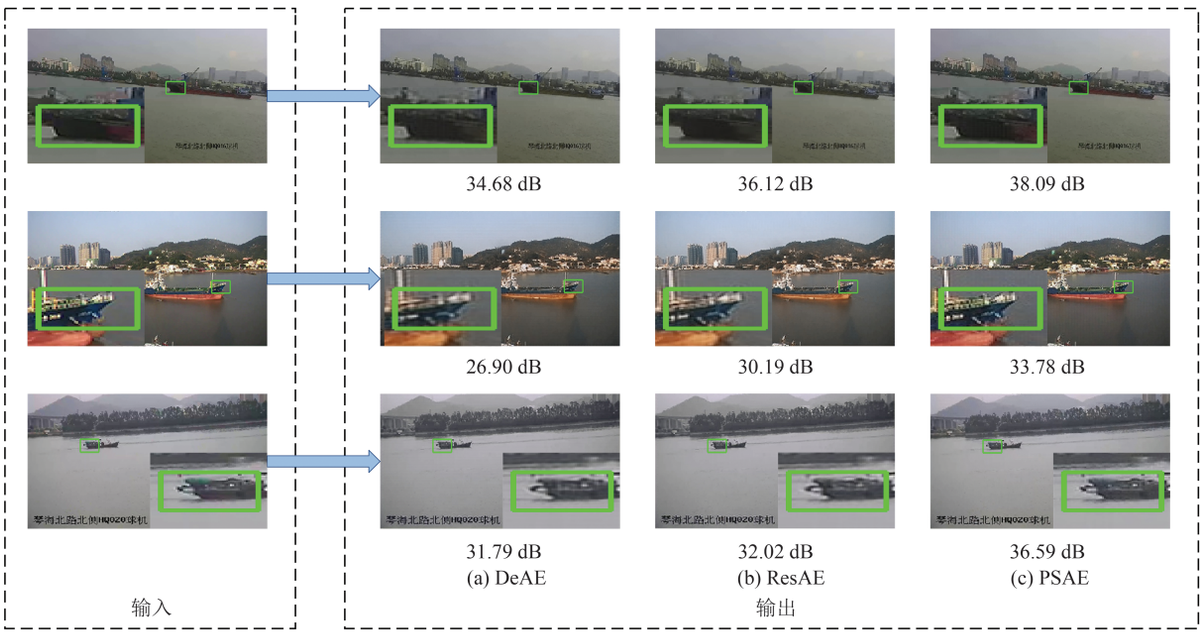

图9　不同模型对图片的重构结果

Fig. 9　Image Reconstruction Results of Different Models

为了将本文提出的空间自编码器模型的性能定量化表示,将图9中的3种模型分别在SAED、Avenue[11]和UCSD[16]3个数据集上进行测试,分别计算出平均PSNR值,如表4所示。由表4可知,本节提出的空间自编码网络模型PSAE在图像重构方面性能最佳。

表4　3种模型在不同数据集上的平均PSNR值/dB

Tab. 4　Average PSNR Values of Three Models on Different Datasets/dB

数据集	DeAE	ResAE	PSAE
SAED	30.03	33.12	36.28
Avenue[11]	30.20	33.56	36.73
UCSD ped1[16]	29.31	32.18	34.27
UCSD ped2[16]	30.05	32.59	35.89

3.7　方差注意力模块对模型性能的影响分析

在本节中,将研究基于方差的注意力模块(variance attention module,VAM)对模型性能的提升效果,实验结果由表5所示。由表5可知,引入方差注意力模块能使模型的检测精度提升2%左右。

在正常事件场景中,视频序列的变化相对较小,因此每个位置的注意力权重是一致的。在异常事件示例中,由于船只比视频中的其他区域具有更快的移动,因此方差注意力模块对船只产生了更高的注意力权重。从相应的注意力特征图中可以看出,快速移动区域的值要远远高于其他区域的值。由于方差注意力模块可以自动为视频片段的运动部分分配更重的权重,因此针对突然快速移动的异常事件会有更高的运动损失,对于异常物体会得到更大的损失,这有助于检测出异常事件。

表5　基于方差的注意力机制对模型性能的影响

Tab. 5　The Impact of Variance Attention Module on Model Performance

方法	AUC/%
本文方法	80.5
本文方法+VAM	82.4
文献[23]	59.2
文献[23]+VAM	61.3

4　结　语

本文提出了一种基于深度时空自编码网络以及多示例学习的视频异常事件检测算法。首先,深度时空自编码网络能提取出视频的RGB特征以及运动特征,该特征在异常事件检测方面有很好的性能,能显著提升检测性能。其次,本文进一步提出了将运动增强网络和多示例学习相结合来检测异常事件的方法,无需对训练视频进

行标注,提高了算法的效率和准确率。本文提出的运动感知特征和多示例排名模型,在船只异常事件视频数据集上取得了很好的效果。未来希望在光线变化的场景检测方面加以突破,使得提出的网络具有更强的鲁棒性,进一步提高检测异常事件的精度。

参 考 文 献

[1] Ma Wenyao. Conformal Detection of Anomalous Behaviors of Vessel[D]. Dalian: Dalian Maritime University, 2018. (马文耀. 船舶异常行为的一致性检测[D]. 大连:大连海事大学, 2018.)

[2] Durso F T, Nickerson R S. Handbook of Applied Cognition[M]. New York: John Wiley & Sons, 2007.

[3] Hu Bo, Liu Ming. Application of Remote Sensing Technology in Situation Awareness of Sea Area[J]. *World Affairs*, 2021(19): 74. (胡波, 刘明. 遥感技术在海域态势感知中的应用[J]. 世界知识, 2021(19): 74.)

[4] Chintoan-Uta M, Silva J R. Global Maritime Domain Awareness: A Sustainable Development Perspective[J]. *WMU Journal of Maritime Affairs*, 2017, 16(1): 37-52.

[5] Shi Z W, Yu X R, Jiang Z G, et al. Ship Detection in High-Resolution Optical Imagery Based on Anomaly Detector and Local Shape Feature[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, 52(8): 4511-4523.

[6] Riveiro M, Pallotta G, Vespe M. Maritime Anomaly Detection: A Review[J]. *WIREs Data Mining and Knowledge Discovery*, 2018, 8(5): e1266.

[7] Liu W, Luo W X, Lian D Z, et al. Future Frame Prediction for Anomaly Detection: A New Baseline[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.

[8] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, New York, USA, 2015.

[9] Hasan M, Choi J, Neumann J, et al. Learning Temporal Regularity in Video Sequences[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[10] Sabokrou M, Fayyaz M, Fathy M, et al. Deep-Anomaly: Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes[J]. *Computer Vision and Image Understanding*, 2018, 172: 88-97.

[11] Lu C W, Shi J P, Jia J Y. Abnormal Event Detection at 150 FPS in MATLAB[C]//IEEE International Conference on Computer Vision, Sydney, Australia, 2013.

[12] Wu P, Liu J, Li M M, et al. Fast Sparse Coding Networks for Anomaly Detection in Videos[J]. *Pattern Recognition*, 2020, 107: 107515.

[13] Stewart R, Ermon S. Label-Free Supervision of Neural Networks with Physics and Domain Knowledge[C]//The 31st AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2017.

[14] Srivastava N, Mansimov E, Salakhutdinov R. Unsupervised Learning of Video Representations Using LSTMS[C]//The 32nd International Conference on International Conference on Machine Learning, Lille, France, 2015.

[15] Simonyan K, Zisserman A. Two-stream Convolutional Networks for Action Recognition in Videos[EB/OL]. (2014-06-09)[2022-01-08]. http://arxiv.org/abs/1406.2199.

[16] Mahadevan V, Li W X, Bhalodia V, et al. Anomaly Detection in Crowded Scenes[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010.

[17] Xu D, Yan Y, Ricci E, et al. Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion[J]. *Computer Vision and Image Understanding*, 2017, 156: 117-127.

[18] Nguyen T N, Meunier J. Anomaly Detection in Video Sequence with Appearance-Motion Correspondence[C]//IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 2019.

[19] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]//IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015.

[20] Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017.

[21] Xie S N, Sun C, Huang J, et al. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification[C]//The 15th European Conference, Munich, Germany, 2018.

[22] Sun D Q, Yang X D, Liu M Y, et al. PWC-Net: CNNS for Optical Flow Using Pyramid, Warping, and Cost Volume[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt

Lake City, USA, 2018.

[23] Sultani W, Chen C, Shah M. Real-World Anomaly Detection in Surveillance Videos[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.

[24] Yang Xianbin, Dang Jianwu, Wang Song, et al. Anomaly Event Detection Based on Two-Stream Network and Multi-instance Learning[J]. *Laser & Optoelectronics Progress*, 2021, 58(20): 2015006. (杨先斌, 党建武, 王松, 等. 基于双流网络与多示例学习的异常事件检测[J]. 激光与光电子学进展, 2021, 58(20): 2015006.)

[25] Zweig M H, Campbell G. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine[J]. *Clinical Chemistry*, 1993, 39(4): 561-577.

[26] Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library [EB/OL]. (2019-12-03) [2022-01-08]. http://arxiv.org/abs/1912.01703.

[27] Stephen K, Menon V. Re Learning Memory Guided Normality for Anomaly Detection[EB/OL]. (2021-04-01)[2022-01-08]. http://arxiv.org/abs/2101.12382.

(上接第 1108 页)

[27] Riedmiller M, Hafner R, Lampe T, et al. Learning by Playing-Solving Sparse Reward Tasks from Scratch[C]//International Conference on Machine Learning, Stockholm, Sweden, 2018.

[28] Volodymyr M, Koray K, David S, et al. Human-Level Control Through Deep Reinforcement Learning [J]. *Nature*, 2015, 518: 529-533.

[29] Bengio Y, Louradour J, Collobert R, et al. Curriculum Learning [C]//The 26th Annual International Conference on Machine Learning, Montreal Quebec, Canada, 2009.

[30] Hussein A, Gaber M M, Elyan E, et al. Imitation Learning: A Survey of Learning Methods[J]. *ACM Computing Surveys*, 2017, 50(2): 1-35.

[31] Morad S D, Mecca R, Poudel R P K, et al. Embodied Visual Navigation with Automatic Curriculum Learning in Real Environments[J]. *IEEE Robotics and Automation Letters*, 2021, 6(2): 683-690.

[32] Fang Q, Xu X, Wang X T, et al. Target-Driven Visual Navigation in Indoor Scenes Using Reinforcement Learning and Imitation Learning[J]. *CAAI Transactions on Intelligence Technology*, 2022, 7(2): 167-176.

[33] Sutton R S, Barto A G. Reinforcement Learning: An Introduction [M]//. Cambridge: MIT Press, 2018.