南京大學

研 究 生 毕 业 论 文

（申请硕士学位）

论 文 题 目 基于深度学习的视频异常检测研究

作 者 姓 名 邵玥

学科、专业方向 计算机技术

研 究 方 向 计算机视觉

指 导 教 师 申富饶 教授

2021 年 **6** 月 **1** 日

学　　号：**MP1833023**

论文答辩日期：**2021** 年 **5** 月 **20** 日

指 导 教 师：　　（签字）

A Research on Video Anomaly Detection Based on Deep Learning

by

Shao Yue

Supervised by

Professor Furao Shen

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
MASTER
in
Computer Technology

Department of Computer Science and Technology
Nanjing University

May 10, 21

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于深度学习的视频异常检测研究

计算机技术 专业 2018 级硕士生姓名：邵玥

指导教师（姓名、职称）：申富饶 教授

摘　要

随着经济的发展和技术的进步，不论是公共场所还是私人区域，都有大量的监控摄像头投入使用。相关人员可以实时查看各个摄像头的监控画面，并对画面作出判断和分析，在有异常情况发生时进行相应的处理。与此同时，基于计算机视觉的视频异常检测技术也发展起来。视频异常检测技术可以通过对视频画面的分析获得相应的异常分数，该分数反映了异常发生的概率，当分数过高时可向相关人员发出警告。使用视频异常检测技术有更强的时效性，较高的报警准确率，能够节约人力资源，并有效避免人员和环境带来的不确定性。

随着对人工智能和深度神经网络研究的不断深入，基于深度学习的视频异常检测技术也有了长足的进步。这种基于深度学习的视频异常检测技术通常通过计算预测或重构误差来判断是否有异常发生。当异常的面积过小时，上述方法容易产生漏报，当遇到难以预测的正常情况，或重构、预测结果有噪声时，上述方法容易产生误报。当前的预测或重构模块主要关注视频帧的外观和运动信息，很少对视频帧的内容进行研究，也没有对视频帧的局部信息进行约束。然而当前视频异常检测领域的主流研究方向是如何更好地重构或预测视频帧，并没有对上述问题进行详细的探讨。

在此基础上，本文引入了一种新的异常分数计算方法，能够对局部的预测或重构误差进行定量分析，并提出了时间一致性的概念，以视频帧的前后信息为参考，对当前帧的异常情况进行判断，缓解了漏报和误报问题。除此之外，本文还提出了内容损失和误差密度损失，在训练网络时对图像的内容和局部误差进行约束，提高了视频异常检测的准确率。本文的主要研究内容和贡献如下：

1. 本文提出了一种基于误差密度和时间一致性的异常分数计算方法。基于误差密度的异常分数计算方法借鉴了卷积神经网络中的平均池化操作，

可以对视频图像分块分析，将每一块的差异进行量化，能够刻画小面积的异常，有效避免了小面积异常漏报的问题。基于时间一致性的异常分数计算方法克服了现有异常分数计算方法只考虑当前帧的缺点，综合考虑前后帧的异常情况后再对当前帧的异常概率进行计算，能够避免正常情况的不可预测性问题和监控视频画面突变问题带来的不良影响，也对噪音更加鲁棒。该方法是一种即插即用的方法，可以与现有方法相结合，直接作用在异常分数的计算阶段。将本文提出的异常分数计算方法与其他方法结合进行实验证明，该方法能够显著提高现有算法在视频异常检测上的性能。

2. 本文提出了一种基于内容损失和误差密度损失的损失函数。内容损失对视频帧的内容进行提取，利用预训练好的模型获得图像的内容特征。使用该损失可以在训练过程中对视频的内容进行约束，提高生成帧的质量。误差密度损失利用误差密度对视频帧的小面积差异进行度量，可以使生成帧与原帧在小面积范围内更接近。实验证明，这两种损失函数均能提高模型的检测水平，且与基于误差密度和时间一致性的异常分数计算方法结合后能够获得更好的效果。
3. 本文将提出来的方法应用到煤矿智能视频分析系统中。该算法对正常情况下的煤矿皮带场景进行学习，获得异常检测的模型。在检测阶段，当皮带上出现异物或者较大煤块时，模型能够得到较高的异常分数，及时发出告警，获得令人满意的结果。

相关实验表明，本文提出的方法可以有效解决视频异常检测中遇到的上述问题，并能显著提高检测结果的准确率。在相关实践中，本文的方法也能得到比较好的表现，具有较高的实用价值。

关键词： 视频异常检测；深度学习；迁移学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS: A Research on Video Anomaly Detection Based on Deep Learning

SPECIALIZATION: Computer Technology

POSTGRADUATE: Shao Yue

MENTOR: Professor Furao Shen

Abstract

With the development of economy and technology, both public and private areas have been using cameras for video surveillance. The relevant staff can view the monitoring images of each camera in real time, make judgments and analysis of the images, and take corresponding treatments when abnormal situations occur. At the same time, video anomaly detection technology based on computer vision has also been developed. Video anomaly detection technology can obtain corresponding anomaly scores by analyzing the video images. The scores reflect the probability of occurrence of anomalies. When the scores are too high, a warning can be issued to relevant staff. The use of video anomaly detection technology is timeliness, can improve the accuracy of the alarm, save human resources, and effectively avoid the uncertainty caused by the staff and the environment.

With the continuous in-depth research of artificial intelligence and deep neural networks, video anomaly detection based on deep learning has also made great progress. This deep learning-based video anomaly detection technology usually judges whether an anomaly occurs by calculating prediction or reconstruction errors. When the area of the anomaly is too small, it tends to cause false negatives. When encountering normal conditions that are difficult to predict, or there are noises in the reconstruction or prediction results, the above methods are prone to false alarms. The current prediction or reconstruction module mainly focuses on the appearance and motion information of the video frame, and rarely studies the content of the video frame, or restricts the local information of the video frame. However, the current mainstream research direction in this field is how to better reconstruct or predict video frames, and the above-mentioned problems have not been discussed in detail.

On this basis, this paper introduces a new anomaly score calculation method, which can quantitatively analyze the local prediction or reconstruction error, and proposes the concept of time consistency, which takes multiple frames before and after the current frame as a reference to alleviate the problem of false negatives and false positives when judging current frame. In addition, this paper also proposes content loss and error density loss, which constrains the content and local errors of the image during network training and improves the accuracy of video anomaly detection. The main research contents and contributions of this paper are as follows:

1. This paper proposes an anomaly score calculation method based on error density and time consistency. The error density-based anomaly score calculation method draws on the average pooling operation in the convolutional neural network, which can quantify the difference of each block in the graph, and can characterize small-area anomalies, effectively to avoid the underreporting of small-area anomalies. The anomaly score calculation method based on time consistency overcomes the shortcomings of the existing anomaly score calculation method that only considers the current frame. After comprehensively considering the abnormal conditions of the previous and subsequent frames, the abnormal probability of the current frame is calculated, which can avoid the bad influence of unpredictability of the normal situation and the sudden change of the surveillance video frame. It is also more robust to noise. This method is a plug-and-play method that can be combined with existing methods to directly act in the calculation phase of anomaly scores. Combining the anomaly score calculation method proposed in this paper with other methods, experiments show that this method can significantly improve the accuracy of the existing algorithms in the detection of video abnormalities.
2. This paper proposes a loss function based on content loss and error density loss. Content loss extracts the content of the video frame and uses the pre-trained model to obtain the features of the image. By restricting the content of the video during the training process, the quality of the generated frame can be improved. Error density loss uses error density to measure the small area difference of video frames, which can make the generated frame closer to the original frame in a small area. Experiments have proved that these two loss functions can improve the detection level of the model and can obtain better results when combined with anomaly score

calculation method based on error density and time consistency.

3. This paper apples the proposed method to the coal mining intelligent video analysis system. This method learns the coal mining belt scene under normal conditions and obtains an anomaly detection model. In the detection stage, when foreign objects or large coal lump appear on the belt, the model can get a higher abnormal score, send an alarm in time, and obtain a satisfactory result.

Related experiments show that the method proposed in this paper can effectively solve the above-mentioned problems encountered in video anomaly detection and can significantly improve the accuracy of the detection results. In related practice, the method in this article can also get good performance and has high practical value.

keywords: Video Anomaly Detection, Deep Learning, Transfer Learning

目　　次

插图清单

附表清单

第一章　绪论

1.1　研究背景及意义

1.1.1　研究背景

随着技术的发展和经济的增长，越来越多的场所开始使用摄像头对现场进行监控和录像，并对记录下来的视频进行存储和分析。在许多公共场所，如地铁、银行、商场、车站等地，相关工作人员可以通过监控系统直观地掌握现场情况，在发生异常事件时及时采取措施，也可以利用录像的回放对事件进行分析和取证，方便对事件的后续处理。在普通居民家里，基于视频监控的家庭安防系统也开始流行。这种监控系统通常可以通过互联网实时传输画面，方便居民在外出时对家里情况进行查看，尤其适合有老人，小孩或者宠物的家庭。传统视频监测技术通常只具有数据采集和存储的功能，主要依靠人工对视频进行实时监控和异常检测。这种方式会消耗大量的人力资源，且不可避免的会产生异常事件的漏报和误报。视频异常检测技术可以对固定角度的监控视频进行分析，检测是否有异常情况发生。因此，通过视频异常检测技术分析监控视频，节约人力资源，提高异常检测的效率和准确率，成为一个具有重要意义的研究课题。

1.1.2　应用方向

不论是在公共场所还是私人场所，视频监控都得到了普遍应用，在此基础上，视频异常检测技术也发挥了巨大作用。截至 2018 年，“中国天网”在大街小巷布置摄像头超过两千万台，是世界上规模最大的视频监控网络，也成为了城市治安的坚强后盾。不仅如此，在火车站，汽车站，景区，道路等人流密集场所和银行，学校等重点安全防范单位里也部署着大量的监控摄像头（见图 1-1[①]）。这些摄像头每时每刻都在产生海量的监控视频数据，其中不乏犯罪，拥堵，非法聚集等异常行为。传统视频监控系统通常需要相关工作人员远

[①]图片来源：http://www.cq.xinhuanet.com/2021-04/26/c_1127376528.htm

程观看监控摄像头采集的视频，并在发生异常时及时告知现场人员采取措施。而视频异常检测技术可以自动分析监控视频，并通过建立能够自动报警的智能监控系统，显著提高异常检测效率和公共场所的安全性。

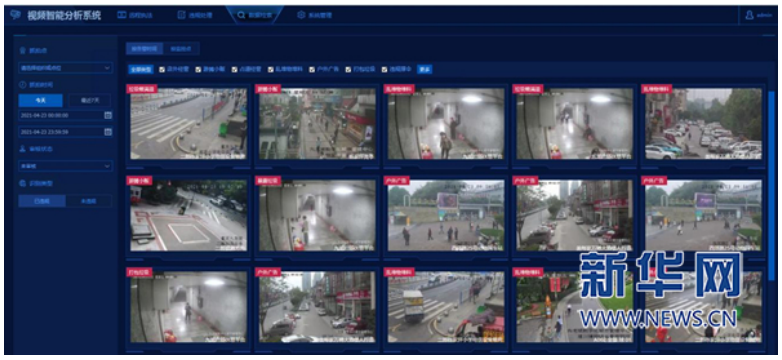


图 1-1: 重庆九龙坡区地下通道实现视频监控全覆盖

企业生产场所可以通过建设监控中心，达到安全生产，质量监督和规范行为等目的。在生产场所使用摄像头对生产线进行拍摄，可以提高生产安全性，避免故障对员工人身安全的威胁。企业还可以利用全厂区监控视频的联动，提高仓储中心的防火防爆水平。来料/成品检验、生产、包装和运输是生产过程的核心环节，利用监控系统对上述环节进行监督，能够及时发现和处理异常，提高产品出厂质量。因此，在生产场所使用视频异常检测技术，对可能出现的异常发出警报，具有非常重要的现实意义。

随着居民安全意识的提高和互联网技术的发展，基于视频监控的家用安防系统得到了普及。这种系统通常可以利用互联网实时传输数据，将家庭监控画面传输到居民面前，方便居民在离家时远程查看家庭内部情况。当有陌生人员从门窗闯入或在室内移动时，视频异常检测系统可以产生警报，提醒居民异常情况的发生，并记录下监控画面，方便居民对突发情况进行后续处理。这种家用监控系统在有小朋友，老人或者宠物的家庭有很大的市场（见图 1-2[①]）。

1.1.3 困难与挑战

视频异常检测问题虽然在近几年已经在学术界引起了越来越多的注意，但是也出现了一些新的挑战。

异常具有环境依赖性，异常的发生与环境之间有很强的关联。对于某些行为，只有发生在特定的场景下才算是异常行为，否则是正常的。行人在路边的

[①]图片来源：https://www.thehour.cn/news/391028.html

图 1-2: 老人通过监控与家人隔空聊天

人行道上行走是正常行为，但是当行人需要穿过马路时，若直接横穿马路是异常行为，若在绿灯时走斑马线就是正常行为。因此，如何在不同的环境下对异常进行准确判断，成了一个研究的难题。

异常具有非平衡性。异常事件的发生概率通常比较小，在视频数据中所占的比重很小，样本量也严重不足。训练样本的缺乏导致如果使用传统机器学习方法进行处理，会产生一系列的问题，并且学习的效果不好，极易欠拟合。

异常具有多样性与未知性。视频中的异常是多种多样的，通常无法穷举，且面临着许多未知的情况。异常的多样性使得视频异常检测算法很难对异常行为进行建模。因此使用传统的分类方法在异常检测领域是不可行的，只能通过对正常行为建模来确定测试数据中是否发生异常。异常的未知性要求视频异常检测算法不仅要对见过的异常进行检测，还要能检测出未来可能发生、但是实际上并没有见过的异常。

最后一个挑战是大规模训练数据的缺乏。当前通用的数据集通常只包含少数场景下以固定角度的摄像机拍摄的视频。最近两年随着视频数据积累，也有一些新数据集被提出[11,12]，但是视频数据的质量还没有得到学术界的认可。目前学术界通用的，仍是一些比较小，场景不多，事件不复杂的数据集。

1.2 研究现状

1.2.1 视频异常检测

根据数据否有异常标签，视频异常检测算法主要分为三类：有监督的视频异常检测算法、无监督的视频异常检测算法和半监督的视频异常检测算法。由

于视频异常检测数据标签的缺乏，本研究主要关注无监督的视频异常检测领域，不对有监督和半监督的视频异常检测进行深入探讨。

无监督视频异常检测技术通过在单一场景的训练视频上学习，对正常情况下的场景进行建模，然后利用获得的模型对测试视频进行检测，判断是否有异常发生。训练视频中只包含正常的情况，因此，当测试视频中的某段场景与训练视频有极大差异时，就视为异常。给定一些不含异常事件的训练视频 $X_{train} \in R^{N_{train} \times r \times c}$，视频异常检测算法的目的是通过对 X_{train} 的学习，获得一个模型，在输入 $X_{test} \in R^{N_{test} \times r \times c}$ 的情况下，能够对 X_{test} 中的每一帧输出一个异常分数[13]。其中 N_{train} 和 N_{test} 分别是训练视频和测试视频的总帧长，$r \times c$ 是视频的分辨率。异常分数是对帧异常水平的定量表示，异常分数越高，该帧为异常的可能性越大。

1.2.2 常见方法

无监督视频异常检测的主流算法分为两类，一类是传统的视频异常检测算法，一类是基于深度学习的视频异常检测算法。

其中，基于传统的视频异常检测算法先对视频进行特征提取，然后对提取到的特征进行异常检测，获得对应的异常分数，若某一帧对应特征的异常分数较高，则视为异常发生。其中，特征提取的好坏对最终检测结果有着至关重要的影响。但是传统视频异常检测算法的特征提取方式通常是手工设计的，特征的好坏强烈依赖于视频的场景内容和设计者的经验水平。

近年来，随着深度学习的不断发展，计算设备算力水平的提升和数据规模的扩大，在计算机视觉领域，基于深度学习的视频目标检测、视频预测、视频表示学习等研究取得了很大进展。在此基础上，基于深度学习的视频异常检测也成为了研究者们关注的课题。这类算法主要分为两种，分别是基于重建误差的视频异常检测算法和基于预测误差的视频异常检测算法。基于重建误差的视频异常检测技术利用自编码器[14] 等网络对视频帧进行压缩和重构。由于训练时的视频数据只包含正常事件，自编码器等模型对异常事件的学习不够充分，异常帧在压缩和重构时的重构误差较大。因此，当重构的帧与原帧有很大不同时，就视为异常情况发生。随着视频预测技术的进步，基于预测误差的视频异常检测技术也逐渐成为主流研究方向。异常情况通常不可预测，因此在使用预测网络对视频进行预测时，若发生异常，预测帧与实际帧差异较大。该方法利用预测帧与原帧之间的差异来确定异常分数的取值，差异大时异常分数更高。

除此之外，还有二者相结合的算法以及一些其他的算法。

1.3 研究内容

本文针对现有基于深度学习的异常检测算法的局限性，如正常事件的不可预测问题，画面的急剧变化问题和小面积异常的漏检问题进行了深入的探讨，提出了一种基于误差密度和时间一致性的视频异常分数计算方法，并对视频异常检测算法的损失函数进行了研究，使用内容损失和误差密度损失对训练过程进行约束。其中主要的研究内容包含如下几点：

本文提出一种基于误差密度和时间一致性的异常分数计算方法。基于误差密度的异常分数计算方法参考了卷积神经网络的平均池化思想，对两张图片之间的误差密度进行估计。当两张图片之间存在误差密度较大的点时，即可获得响应。利用误差密度对视频帧的异常程度进行估计，能够显著降低小面积异常的漏检率。基于时间一致性的视频异常计算方法将视频中前后帧对当前帧的影响反映到异常分数的计算过程中，能有效地去除噪声点对异常判断的影响，避免正常事件的不可预测性和画面急剧变化带来的误检问题。本文设计实验对上述方法进行验证，证明该方法能够有效提高异常检测的表现，降低漏检和误检率。

本文提出一种基于内容损失和误差密度损失的损失函数设计。在训练视频异常检测的神经网络时，常用的损失通常只考虑了视频中图像的外观和视频中物体的描述，而不是视频中具体的内容。基于内容损失的视频异常检测算法利用预训练的深度神经网络提取视频中的深度特征，可对视频中每一帧的图像进行内容上的约束，保证网络训练过程中生成帧和预测帧内容上相近。误差密度损失使用误差密度对图像进行约束，保证两张图像没有局部差异大的地方。实验证明，内容损失和误差密度损失能够在训练阶段发挥约束作用，提高生成帧的质量，在与基于误差密度和时间一致性的异常分数计算方法结合时也能显著提高算法表现。

在上述工作的基础上，本文将该方法应用于煤矿视频智能分析系统，对煤矿皮带负载情况进行异常检测。当监控视频中的皮带上有异物或者大煤块出现时，系统能够产生报警，提醒相关工作人员进行处理。

1.4 论文纲要

本文主要研究了基于深度神经网络的视频异常检测算法，并提出了两种改进方案，一种是基于误差密度和时间一致性的视频异常分数计算方法，一种是基于内容损失和误差密度损失的损失函数设计，本文设计实验验证了上述算法的有效性，并将该算法应用于煤矿智能视频分析系统。本文共分为六章，第一章为绪论，简要介绍了视频异常检测问题的研究背景和意义，并对相关研究现状进行了简要的总结；第二章为相关工作，对基于传统的视频异常检测算法和基于深度学习的视频异常检测算法进行了梳理，并对其中几个经典的研究内容进行了介绍；第三章介绍了基于误差密度和时间一致性的异常分数计算方法并对其进行实验验证；第四章介绍了基于内容损失和误差密度损失的损失函数设计并对其进行实验验证；第五章将算法应用与煤矿智能视频分析系统；第六章对全文进行总结和对未来工作进行展望。

第二章　相关工作

随着监控摄像头的普及，视频异常检测问题的研究也得到了越来越多的关注。早期视频异常检测主要是对视频进行处理，获得相应的特征，然后对特征进行异常检测。近年来，随着深度学习，尤其是深度神经网络在各个领域的兴起，基于深度学习的视频异常检测方法也逐渐成为学者研究的方向。本章先简要介绍了视频异常检测的常见数据集和评价指标，然后介绍了当前的主流方法，并对其中具有代表性的研究工作进行介绍。

2.1　数据集

视频异常检测的数据集通常来自于监控摄像头，视频的背景是固定的，需要对前景中的物体进行异常的检测。下面对一些常见的数据集进行简要介绍：

UCSD Pedestrain 数据集[15] 是一个被广泛使用的数据集。该数据集分为两个部分，分别为 Ped1 和 Ped2。这两个数据集分别采集自两个不同的监控摄像头，记录了人行道上的场景。Ped1 数据集有 36 个测试视频，Ped2 数据集有 12 个测试视频，两个测试集共有 5 种不同类型的异常，已经被标注出来，标签分别为：骑自行车，滑滑板，小型机动车，横穿草坪和其他。

CUHK Avenue 数据集[1] 采集自一个固定在户外的监控摄像头，该摄像头对一栋建筑物外的人行道进行了记录。数据集包含 16 个训练视频和 21 个测试视频，拍摄的是行人走进走出该栋建筑物的场景。测试视频里的异常行为有：人朝空中扔纸和书包，奔跑，跳舞等，但只标注了是否发生异常，未对具体的异常类型进行标注。图 2-1展示了两个异常行为，左图展示的是奇怪动作的异常，其中有人在人群中奔跑；右图展示的是行人走错方向的异常。 Avenue 数据集包含的挑战有：测试视频中含有轻微的摄像头抖动；训练视频中有少量的异常行为；一些测试视频中的正常行为在测试数据中很少出现。

Subway 数据集[16] 由两个长视频组成，分别是 entrance 和 exit，拍摄地铁的入口处和出口处。异常行为包含人走错方向，清洁墙壁等。上述数据集都是针对单一场景的，除此之外还有一些多场景的数据集，其中比较著名的数据集

图 2-1: Avenue 数据集的异常行为[1]

表 2-1: 单一场景数据集

数据集	总帧数	异常事件	异常类型	分辨率
UCSD Ped1	14,000	54	5	238 × 158
UCSD Ped2	4560	23	5	360 × 240
Subway entrance	86,535	66	5	512 × 384
Subway exit	38,940	19	3	512 × 384
Avenue	30,652	47	5	640 × 360

有两个，一个是由上海科技大学提出的 ShanghaiTech 数据集[5]，它由 13 个不同场景的训练和测试视频组成。还有 Sultani 等人提出的 UCF-Crime 数据集[11]，是由网络上收集的很多监控视频组成的，这些视频都有着对应的异常标注。表格 2-1列出了上述数据集中单一场景的数据集。

2.2 性能评价指标

2.2.1 指标

视频异常检测的结果通常根据异常分数来判断，当异常分数高于某一阈值时，将该帧标记为异常，当异常分数低于该阈值时，则被标注为正常。异常检测的各种可能结果见表 2-2。其中真正例表示异常的目标被检测为异常的情况，假正例表示正常的目标被检测为异常，真反例表示正常的目标被检测为正常，假反例表示异常的目标被检测为正常。

在此基础上，我们可以使用真正例率（True Positive Rate，TPR）和假正例率（False Positive Rate，FPR) 来判断异常检测算法的好坏，计算方法见公式 2-1和 2-2。

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2-1}$$

表 2-2: 异常检测的可能结果

真实情况	预测情况	
	正例	反例
正例	真正例（True Positives，TP）	假反例（False Negatives，FN）
反例	假正例（False Positive，FP）	真反例（True Negatives，TN）

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{2-2}$$

由于视频中异常场景的帧数往往远小于正常场景的帧数，具有非平衡性，而上述两种指标会受到这一性质的影响，因此在业界使用的不多。通过对异常分数的阈值进行调整，可以得到不同假正例率下的真正例率。以 FPR 为横坐标，TPR 为纵坐标，可以得到接受者操作特征曲线（Receiver Operating Characteristic curve，ROC 曲线），然后计算出 ROC 曲线下的面积（Area Under Curve，AUC）。AUC 能够有效地衡量视频异常检测算法的优劣（图 2-2），AUC 越高，说明算法区分正常和异常情况的能力越强。

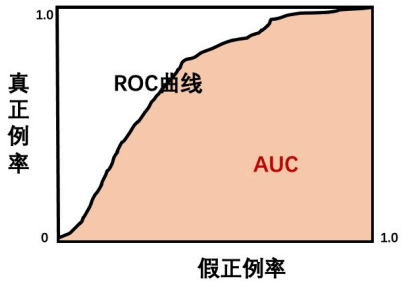


图 2-2: ROC 曲线

2.2.2 衡量标准

衡量异常检测算法常用的标准有两种，一种是基于帧评价方法，一种是基于像素的评价方法。基于帧的评价方法以帧为单位来判断异常，当帧中有某个像素点被视为异常时，整个帧被标注为异常。当视频帧中所有的像素都正常时，整个帧被标注为正常。基于像素的评价方法以像素为单位来判断异常，当帧中 40% 以上的异常像素被检测出时，才视为异常发生。若异常帧中只检出了小于 40% 的异常像素，则该帧被标注为正常，为假反例。

2.3 传统方法

传统的视频异常检测算法先对视频的特征进行提取，然后在得到的特征上使用异常检测算法，判断是否有异常发生。在训练阶段（见图2-3），首先对不包含异常事件的训练视频 X_{train} 进行特征提取，获得训练集上的特征表述 F_{train}，然后通过某种方式对这些特征进行建模，获得相应的模型。在测试阶段(见图2-4)，算法将测试视频 X_{test} 对应的特征 F_{test} 输入到模型中，获得相应的异常分数，判断是否发生异常。

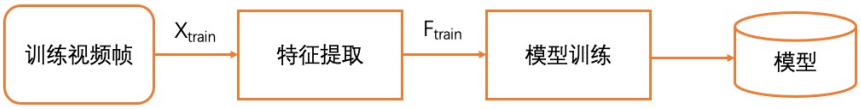


图 2-3: 传统视频异常检测的训练阶段

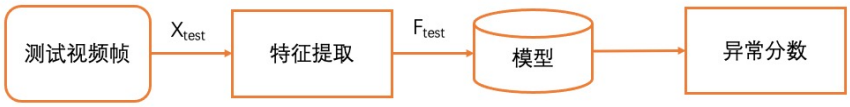


图 2-4: 传统视频异常检测的测试阶段

2.3.1 特征提取

传统的视频异常检测算法首先对视频的特征进行提取，获得能够对视频进行准确描述的特征。特征的好坏会直接影响后续异常的检测，是算法中的核心步骤之一。

方向梯度直方图 (Histogram of Oriented Gradient，HOG)[17] 是一种常用的特征提取方法，它使用梯度的方向密度来描述图像中目标的轮廓，已经被广泛应用于图像领域。HOG 在图形的几何和光学发生形变时具有不变性，可以准确刻画视频的外观特征，方便后续异常的检测。光流直方图 (Histograms of Optical Flow，HOF)[18] 引入了光流（Optical Flow）[19] 信息，能对图像中的运动信息进行表示。图2-5[①]展示了光流计算的一个实例。左一和左二两张图是实际运动的图像，有一些椅子、沙发等家具在空中飘浮，图片左上角的木头椅子一开始

[①]图片来源：https://github.com/ClementPinard/FlowNetPytorch

全在视野之中，后来向上飞去，只能看到剩下的一半。右一是计算得到的光流信息，该木头椅子在光流图的左上角，呈现紫红色。从图中可以看到光流能够克服背景的干扰，对这些家具的运动行为进行准确的刻画。HOF 对光流的方向进行统计，能有效地表示视频的运动情况。HOF 既能像光流一样表达出动作信息，又弥补了光流对尺度和方向敏感的缺点。在此基础上，Yang 等人提出了多尺度光流直方图（multi-scale histogram of optical flow，MHOF）[20]，对不同幅度范围的光流进行统计，能更好的利用光流幅度信息。

图 2-5: 运动的光流信息

除此之外，还有粒子轨迹[21]，社会力模型（Social Force model）[22] 和混合动态纹理（Mixture of Dynamic Textures，MDT）[23] 等特征提取方式。

2.3.2 异常检测

异常检测是指发现数据中不符合预期的样本。图2-6展示了二维数据下的异常数据。其中，N_1 和 N_2 是数据的主要分布区域，因此该区域中的数据被视为正常数据。而区域 O_3，数据点 O_1、O_2 和其他一些远离主要区域的点被视为异常。在传统的视频异常检测算法中，提取特征后会对正常的视频特征进行建模，利用不同的异常检测模型对特征中的异常进行检测。

常见的异常检测算法有基于统计的异常检测算法[24,25]，这类算法假设正常数据服从一定的分布，比如正态分布，高斯混合分布等，而异常数据与正常数据有着不同的分布。算法利用统计学习方法学习分布模型的参数，然后将数据拟合到模型中去，若某个数据点属于这个分布的概率较低，则大概率被判断为异常。这类算法主要有基于高斯混合模型的算法[26]，基于回归模型的算法[27]，基于核密度的方法[28,29] 和一些其他的统计算法[30]。这些统计算法易于理解和实现，但是需要数据满足一定的分布假设。

基于密度的异常检测算法对数据点的密度进行估算，并假设数据分布相对集中，位于低密度区域的点被视为异常。基于局部异常因子（Local Outlier Factor，LOF）的算法[31] 通过点的 k 邻域计算出该点对应的密度，并根据密度

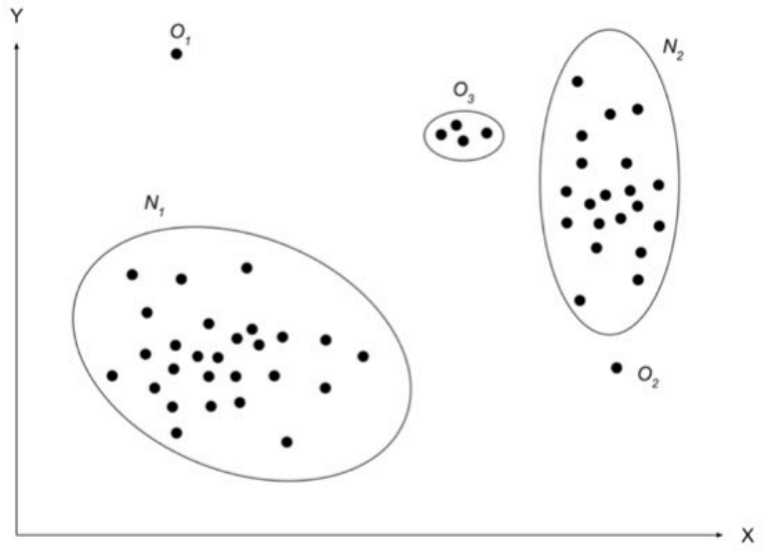


图 2-6: 二维数据下的异常检测[2]

高低判断是否为异常。Tang 等人[32] 提出了基于连接性的异常因子（Connective-based Outlier Factor，COF），利用最短路径方法估算密度，能够对流形数据进行处理。除此之外，还有 INFLO[33]，LoOP[34] 和 RDOS[35] 等算法。此类算法不需要对参数进行计算，也无需对数据分布进行假设，但是通常实现相对复杂，比基于统计的异常检测算法更难计算。

基于距离的异常检测算法通过计算点之间的距离来判断是否为异常。若一个点距离其最近邻很远，那么这个点极有可能是异常。其中最常见的异常检测算法是基于 K 近邻 (K-nearest Neighbor，KNN)[36] 的异常检测算法。Knorr 等人[37] 提出了一种基于距离的算法，若点在距离 d 内有少于 p 个邻居，则视为异常。Ramaswamy 等人[38] 提出的算法则是对所有样本第 k 近邻的距离进行排序，选出其中最大的 n 个，将其标记为异常。Angiulli 等人[39] 提出的算法对所有样本前 k 近邻的平均距离进行排序，选出其中最大的 n 个作为异常输出。基于距离的算法也不需要对数据分布进行假设，并且易于理解。但是当数据维度比较高时，需要进行特殊的处理，否则不论是计算难度还是最终判断结果都会受到影响。

基于聚类的异常检测算法使用聚类的思想对数据是否为异常进行判断。这类方法主要有三种假设，一是把不属于任何聚类的点视为异常，二是把离最近聚类簇较远的点视为异常，三是把稀疏聚类簇和较小聚类簇内的点视为异常。基于聚类的局部异常因子算法 (Clustering-Based Local Outlier Factor，

CBLOF)[40] 是这一类的代表算法之一，它通过计算 CBLOF 来表示数据点属于簇的概率，点的 CBLOF 值越低，被视为异常值的概率越大。

基于树的异常检测算法通过子空间的划分确定是否为异常。这种方法不用计算距离，且能够对任意形状分布的异常点进行划分。Liu[3] 等人发现异常数据通常很少且与正常数据不同，所以这些数据很容易被孤立。当对数据用随机树算法进行划分时，图 2-7(a) 中的正常点 x_i 需要进行多次划分才能被孤立。与之相对的是图 2-7(b) 中的异常点 x_o，只需要进行少数几次的划分即可被孤立。而划分的次数对应着数据点在树中的路径长度，因此，我们可以得出这样的结论：异常点对应的路径比正常点对应的路径更长。由于树的划分是随机生成的，所以作者用多个树构成森林，对路径的平均长度进行统计，并在图 2-7(c) 中展示了点 x_i 和点 x_o 的平均路径长度。可以看出，当使用的树数目越多时，两个点的平均路径长度趋于稳定，且点 x_i 的平均路径长度明显大于点 x_o 的平均路径长度。

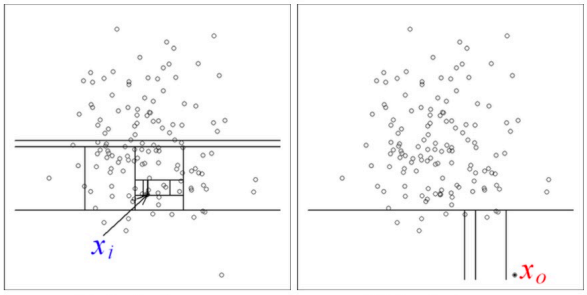
在此基础上，孤立森林（Isolation Forest，IForest) 算法被提出。该算法分为两个阶段，第一个阶段是训练阶段：用数据构造孤立树，并用多个孤立树组成孤立森林；第二个阶段是评估阶段：将数据点带入每棵孤立树，计算该点的平均高度，根据高度对异常分数进行计算。在 IForest 算法之后，还有一些基于树的异常检测算法被提出，并应用于多个异常检测领域[41,42]。

除此之外一些其他的异常检测算法，比如基于集成的异常检测算法[43]，基于学习的异常检测算法[44]，基于单类学习的异常检测算法[45] 等。通过上述算法对训练视频的特征 F_{train} 进行学习后，建立相应的模型，即可利用模型对测试视频的特征 F_{test} 进行判断，获得对应的异常分数。

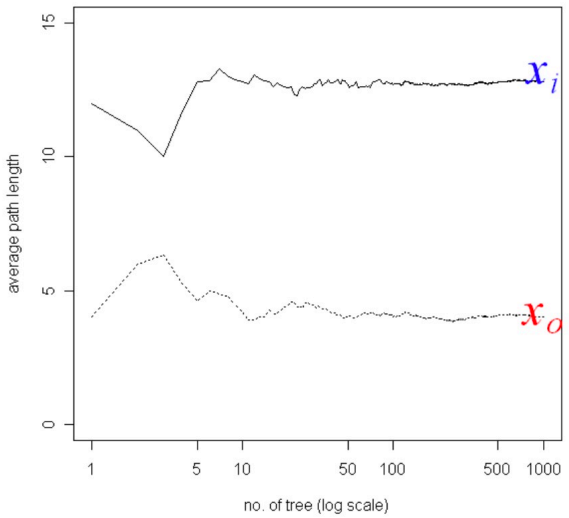
2.3.3 总结

传统的视频异常检测算法通过特征提取，建模，测试等阶段对视频中是否含有异常进行判断，其中较为关键的步骤是特征的提取和模型的建立，这两个环节的好坏直接决定最终算法的优劣。

在特征提取步骤中，使用简单的底层特征计算方便，但不能对监控场景的画面进行很好的描述，最终表现欠佳；使用复杂设计的特征提取方式则能有效的提取外观和运动的信息，但是计算和设计都比较复杂。特征的提取通常与检测的场景有较强的相关性，针对不同的应用场景需要选择不同的特征。如在监控车流的应用中，选取的特征需要对车辆的轨迹进行准确描述，才能判断是否



(a) Isolating x_i (b) Isolating x_o



(c) Average path lengths converge

图 2-7: 异常数据的易孤立性[3]

有闯红灯、超速等异常行为。而在监控人群异常的应用中，则需要对人群进行建模，找到能准确描述人群的特征，这些特征能在发生人群踩踏、骚动等异常时发生明显变化，方便对各种人群异常情况进行捕捉。这些特征需要根据对应场景进行相应的设计，但是通用性欠佳，且要求特征设计者具有一定的专业素养和相关经验，才能提出好的特征提取方式。

在模型建立阶段，也需要针对不同的场景和特征，选择不同的模型。若在某些特定场景下，特征恰好服从某一分布，则用基于统计的异常检测算法能够得到很好的效果，且设计简单，方便理解。但当特征不服从相应的分布时，统计模型不能对相应的特征进行建模，得到的检测结果准确度较低。当特征维数不高时，使用基于相似性的算法进行异常检测通常可得到不错的结果，但是当特征的维数升高时，不论是距离还是密度的计算都会变得困难，需要使用专门针对高维数据设计的异常检测算法来避免这些问题。因此，在模型建立阶段，

也需要针对场景和特征选择不同的模型，模型选择的不好会得到很差的检测结果。

2.4 深度学习方法

近年来随着深度学习的不断发展，深度神经网络已经在图像目标识别、视频生成、图像分类等领域取得了很好的效果。与传统的异常检测算法不同，深度学习能够自动提取出良好的特征，在视频异常检测任务上获得比较好的结果。基于深度学习的异常检测算法主要分为两类，一是基于重构误差的异常检测算法，二是基于预测误差的异常检测算法。除此之外，还有基于重构和预测误差相结合的算法和一些其他算法。

2.4.1 基于重构误差的方法

基于重构误差的异常检测算法将视频帧压缩成一个编码，然后把编码还原成视频帧，根据原帧与重构帧的误差来判断是否为异常。由于在训练阶段只使用了正常视频，模型对异常事件重构的能力比较差，在发生异常时，会产生较大的重构误差。

Hasan 等人[14] 提出了基于自编码器（Auto-Encoder）的异常检测算法。这是一个端到端的学习框架，能够对连续多个视频帧及其特征进行重构，学习正常事件的模型，并利用重构误差判断异常的概率。该框架包含两个自编码器：其中一个是全连接自编码器，用来学习手工提取的特征，包括 HOG，HOF 和轨迹信息；另一个是全卷积自编码器，用来学习多个视频帧。该方法使用两个自编码器的重构误差对异常进行检测，并对相应的误差进行了分析。

Luo 等人[46] 使用卷积长短记忆网络自编码器（Convolutional LSTMs Auto-Encoder，ConvLSTM-AE) 同时对正常视频的外观和运动模式进行建模。ConvLSTM-AE 首先使用两个网络对视频进行处理，一个是卷积神经网络，用来对视频每一帧外观上的特征进行编码；一个是卷积长短记忆网络（Convolutional LSTM，ConvLSTM)，用来记忆过去帧的运动信息。然后将这两个网络与自编码器结合在一起，得到一个能同时学习正常视频外观和运动信息的模型。

Nguyen 等人[4] 提出了一种满足外观和运动一致性的视频异常检测算法。网络的框架如图 2–8所示，该网络的输入是一个视频帧 x_t，共包含三个主要部

分。左侧的部分是普通的编码器，可将视频帧 x_t 进行压缩，获得该图像对应的特征 z_t。右上部分是外观解码器，输入是 z_t，输出是解码后的图像 $\hat{x}_t$。右下部分是运动解码器，输入也是 z_t，但是输出的是对应的光流图像 $\hat{F}_t$。网络的具体结构借鉴了 U-net[47]，在编码器和运动解码器之间加入了一些跳跃连接，可将低层次的特征直接从原始域传递给解码器。但在编码器和外观解码器之间不使用跳跃连接，避免直接传递低层次特征对图像重构的影响。算法的目标是在学习外观信息的同时，对运动信息进行学习，使 $\hat{x}_t$ 与 x_t 之间，$\hat{F}_t$ 与实际光流 F_t 之间更接近。为了达到此目的，算法中引入了生成对抗网络（Generative Adversarial Network，GAN）[48] 对整个网络进行训练，获得了比较好的效果。

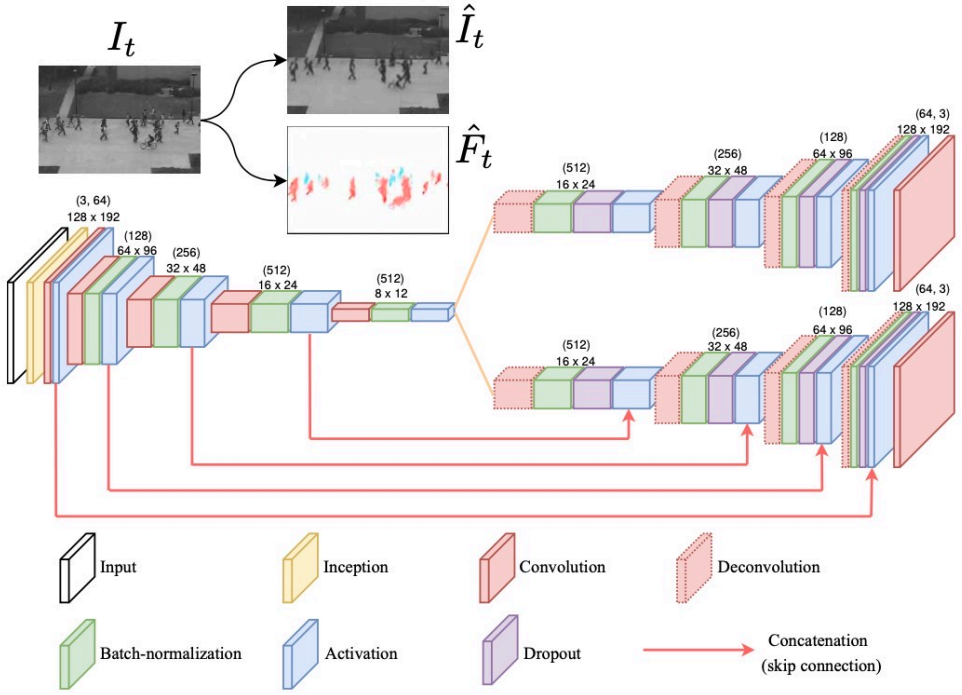


图 2-8: 一种满足外观和运动一致性的视频异常检测算法[4]

Zaheer 等人[49] 提出了一种基于 OGNet（Old is Gold Net）的视频异常检测算法。该算法也使用了 GAN 的思想，对重构过程进行训练。但与其他方法不同，它并不是利用判别器判断输入图片的真假，而是判断重构的图像是否是高质量的。算法还提出了伪异常模块，利用正常的训练数据产生伪异常样本，使得模型能够更好地对正常事件进行刻画，提高了模型的稳定性和准确性。

然而，基于重构的异常视频监测算法有着难以训练、容易过拟合的缺点。算法有时会过好地重构异常事件，使得检测结果准确率不高。

2.4.2 基于预测误差的方法

基于预测误差的异常检测算法通常采用某种视频预测算法对监控视频进行预测，并将具有较大预测误差的帧视为异常帧。Medel 等人[50] 使用一种基于卷积长短记忆网络 (Convolutional LSTM，ConvLSTM) 的预测模型对视频序列进行预测，然后根据误差进行异常检测。一些比较常见的生成模型，比如 GAN 和变分自编码器（Variational Auto-Encoder，VAE）[51] 也被用来做基于预测误差的异常检测。

Liu 等人[5] 提出了一种基于预测的视频异常检测算法（见图 2-9)。该算法的输入为一个视频帧序列，利用 U-Net 对视频序列的下一帧预测。在预测的同时还增加了基于 FlowNet[52] 的光流约束，使得预测帧与原帧之间的运动信息得到保留。在训练过程中，该算法使用了 GAN 的思想，将 U-net 网络作为生成器进行训练，另外使用一个判别器来判断输入的是真实帧还是预测帧。判别器的训练目标是能够更好的分辨真实帧和预测帧，而生成器的目标是能够生成骗过判别器的预测帧，二者进行对抗训练，能够取得较好的训练结果。在测试阶段，算法的输入为测试视频帧，生成器生成预测帧，然后根据预测帧计算预测光流，最终对预测帧、真实帧之间的误差和预测光流、真实光流之间的误差求加权平均，获得对应帧的异常分数。

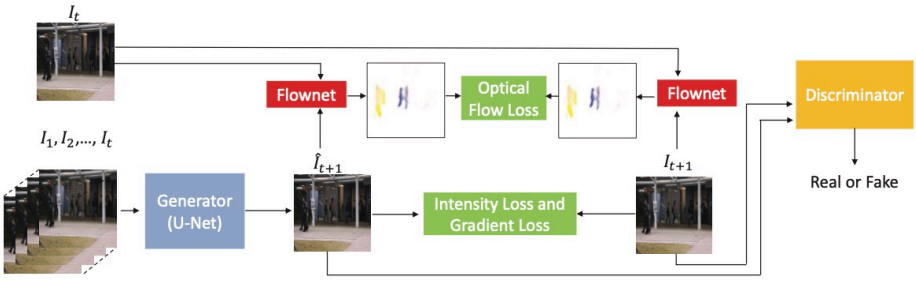


图 2-9: 一种基于未来帧预测的异常检测算法[5]

Rodrigues 等人提出了[53] 一种多时间尺度的算法，该算法主要针对视频中的人类异常行为进行检测。为了更精准地描述人类行为，该算法的输入是人的姿态轨迹，然后对姿态轨迹特征进行向前和向后预测。不同的异常持续的时间可能不同，而当前的主流算法主要针对单一时间尺度进行判断，为了克服这一问题，该算法提出了一种多时间尺度的异常检测算法，首先在不同的时间尺度上对视频进行预测，然后结合不同尺度上的异常检测结果，获得最终的异常分数。

2.4.3 基于重构和预测误差相结合的方法

Morais 等人[6] 提出了一种基于骨架轨迹的异常检测算法（图 2-10），主要对行人的异常行为进行检测。算法对人的特征进行抽取，获得全局身体运动和局部身体运动的表示，然后将这些特征输入到信息传递编码解码循环网络中 (Message-Passing Encoder-Decoder Recurrent Network，MPED-RNN)，进行异常的判断。网络共有两个流，一个流处理全局运动特征，一个流处理局部运动特征，并分别对这两个特征进行重构和预测。

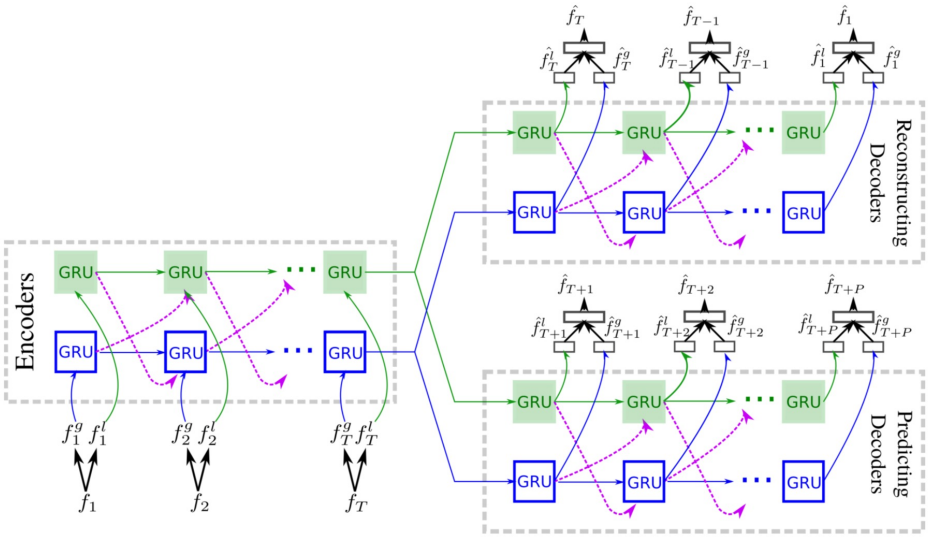


图 2-10: 一种基于人体骨架轨迹的异常检测算法[6]

Zhao 等人[54] 利用三维卷积网络构建自编码器，并在重构的基础上使用预测模块对预测帧进行预测。但该预测结果并不用于异常分数的计算，只是为了让自编码器更好地对时空特征进行提取。

2.4.4 其他方法

Park 等人[7] 在自编码器的基础上加上了记忆模块，用来记录正常模式的原型（图 2-11）。记忆模块具有更新机制，能够对自编码器学习到的隐向量进行记录和更新。记忆模块的使用限制了卷积神经网络的能力，避免对异常重构得过好造成漏检。为了对记忆模块进行训练，算法中还提出了两种新的损失，分别是聚敛损失和分离损失。聚敛损失使获得的特征尽可能的接近其对应项的内容，但单纯使用聚敛损失会使记忆模块中的项没有区分度。在此基础上，作者

又使用了分离损失，使得不同项之间的差异尽可能的大。在测试阶段，算法异常分数的计算公式如下：

$$S_t = \lambda(1 - g(P(\hat{I}_t, I_t)) + (1 - \lambda)g(D(q_t, p))) \tag{2-3}$$

其中 $P(\hat{I}_t, I_t)$ 是基于重构误差计算得到的分数，$g(D(q_t, p))$ 是基于记忆误差计算得到的分数，用来衡量隐向量与记忆模块中项的距离。当视频帧为异常时，其计算得到的隐向量与记忆模块中已有记忆差异较大，而当视频帧不包含异常时，计算得到的隐向量通常与已有记忆比较接近。实验表明，增加了记忆模块以后，算法的准确率得到了一定的提升。

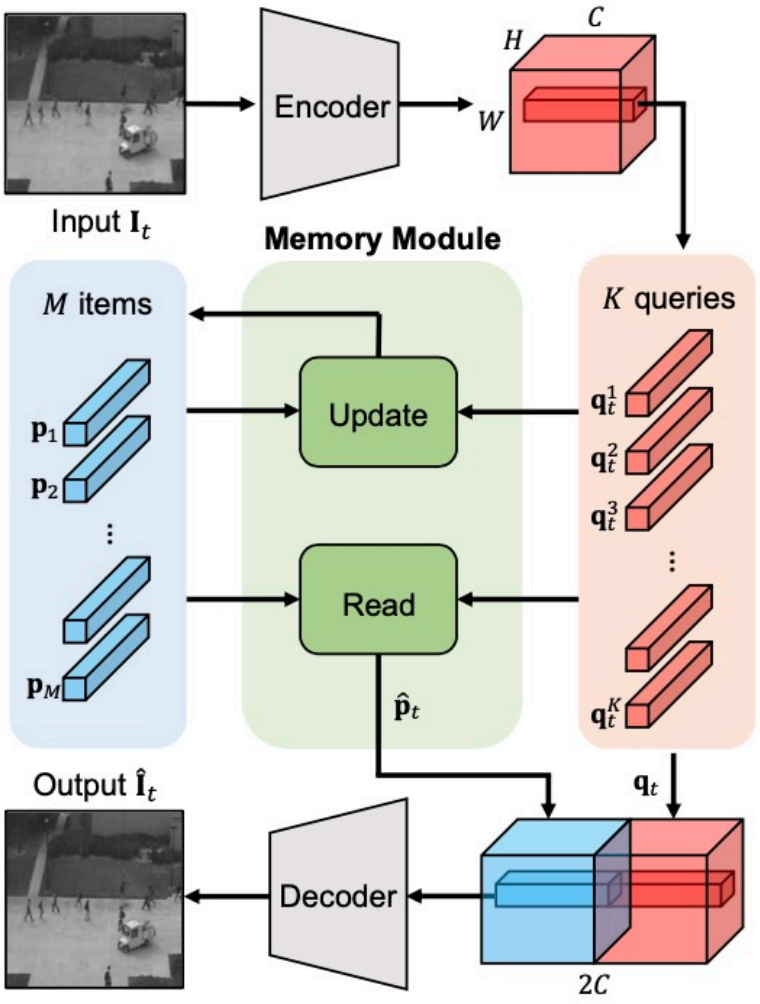


图 2-11: MNAD 算法[7]

Chang 等人[8] 提出了一种全新的卷积自编码器结构来分别学习视频的空间

和时间信息。其中一部分网络负责重建视频帧序列的最后一帧（last individual frame，LIF)，作为视频空间特征的提取，另一部分网络以视频序列为输入，输出 RGB 差值图来模拟光流信息，作为视频时间特征的提取。图 2-12中左图为实际的 RGB 视频帧，中间是帧对应的 RGB 差值，右图是光流图像，可以看出 RGB 差值也可以像光流一样，很好地对视频的运动信息进行表示。在自编码器的训练过程中，该方法还用了一种基于深度 k 均值聚类的方法，使获得的隐向量更为紧凑。

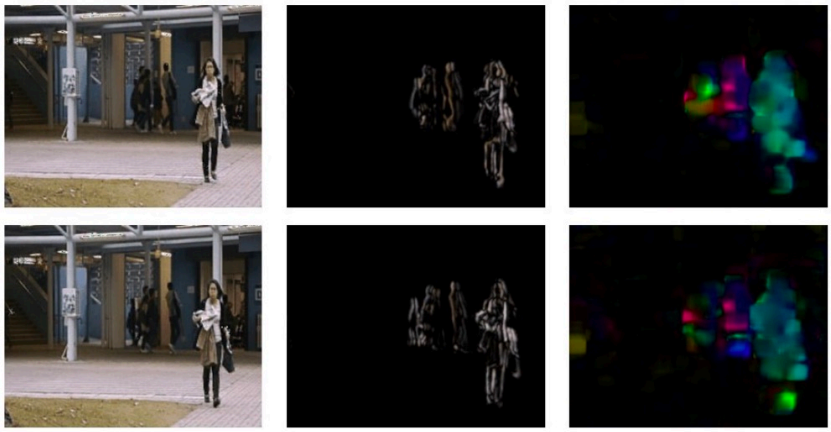


图 2-12: 一些视频帧的图像，RGB 差值图和光流图[8]

Ionescu 等人[55] 利用神经网络进行特征提取，对获得的运动和外观特征使用向量支持机和 k 均值算法进行异常检测。Vu 等人[56] 提出了一种多层信息融合的方法进行视频异常检测，该方法利用不同层次的语义信息进行异常判断，获得了更可靠和准确的结果。Yu 等人[57] 借鉴完形填空的思想，并不是对帧进行前向或者后向预测，而是从帧序列中抽取一帧，训练网络对该帧进行补充，最终根据补充的情况来判断异常。

2.4.5 总结

基于深度学习的异常检测算法利用深度神经网络将特征提取和异常检测步骤结合在一起，通常比传统算法更为高效，可以实现端到端的训练过程。

基于重构的视频异常事件检测算法利用神经网络对视频进行压缩和解码，当解码获得的重构视频与原内容差异较大时，发生异常的可能性也变大。但是随着压缩解码技术的发展，深度神经网络的泛化能力越来越好，甚至能在发生异常时产生高质量的重构结果，导致漏检的发生。基于预测的视频异常检测事

件使用神经网络对视频进行预测，充分利用了异常的多样性和未知性，当预测误差较大时视为异常发生。但在某些情况下，正常事件具有不可预测性，比如闪烁的车灯，突然的开门，变色的交通灯，快速行驶入镜头的车辆等。当上述情况发生时，由于其无法事先预测，得到的预测误差会很大，易发生误检。除此之外，还有一些其他的方法，这些方法通常是上述两种方法的结合，或者与传统的视频异常检测方法相结合，因此，也具有上述方法的缺陷。

这些方法都是基于误差来实现的，而主流的误差计算方法都是对整个画面误差进行求和。当异常发生的范围比较小时，如突然掏出手枪，得到的异常值也比较小，容易发生漏检。当画面发生急剧变化时，如室外摄像机由于雨雪冰雹等外力产生抖动，或者有飞虫突然挡住镜头，产生的误差很大，容易发生误检。

2.5 本章小结

本章主要总结了视频异常检测的一些常用方法，并对他们的优势和局限性进行了讨论。我们可以看出，传统视频异常检测算法和基于深度学习的视频异常检测算法虽然在异常检测问题上已经有比较好的表现，但是也会遇到2.4.5小节中的一些问题。针对上述问题，本文将提出两种视频异常检测方法来缓解这些问题，并进行相关实验，验证方法的有效性。

第三章　基于误差密度和时间一致性的视频异常检测方法

第二章介绍了现有视频异常检测技术和其存在的一些问题，如：正常事件的不可预测问题，画面的急剧变化问题和小面积异常的漏检问题，为了应对这些挑战，本章提出了一个基于时空一致性和误差密度的视频异常分数计算方法。本研究利用异常的时空的一致性缓解不可预测的正常事件和画面变化产生的误检，采用误差密度分析来克服小面积异常的漏检。本章首先给出视频异常检测的问题描述，然后介绍算法的核心思想，最后阐述模型的实验结果，并对结果进行分析。

3.1　现有基于深度学习的异常检测算法的局限性

视频监控的不断普及使得视频异常检测技术成为研究人员关注的热点。传统的人工检测耗费较多的人力资源，可靠性和准确度也不能得到保证。利用算法对视频进行自动地异常检测可以显著地提高效率，也能避免检测者个人原因导致的不稳定性。当前主流视频异常检测算法主要是无监督的形式，利用不含异常的视频进行学习或训练，得到正常的行为模式，然后对测试视频进行分析，判断其是否含有异常。随着深度学习技术的不断发展，计算机视频分析、预测等领域都取得了丰硕的成果，视频异常检测技术在此基础上也获得了很好的效果，但仍有下述问题亟待解决：

首先是正常事件的不可预测性带来的误检问题。基于预测的视频异常检测技术利用异常的未知性来检测异常。未知的异常不易预测，因此预测结果会与原视频帧有极大的差异，可以利用此差异来判断是否有异常发生。但是在日常生活中有很多正常的事件也是不可预测的，比如车灯的突然闪烁，门的突然打开等。当此类事件发生时，虽然不是异常，但是由于预测的效果不好，预测帧与原视频帧有较大的差异，易被误判为异常。

当画面发生急剧变化时，会造成异常值的波动。监控摄像头，尤其是室外

的摄像头，很容易由于不可抗力因素产生画面的急剧变化，比如飞虫接近摄像头时会遮挡画面，或者当摄像头因雨雪天气发生抖动时，画面会随之抖动。当此类情况发生时，不论是基于重构误差还是基于预测误差的视频异常检测算法都会产生极大差异，造成误检。

除此之外，还有小面积异常的漏检问题。当前基于深度学习的异常检测技术主要是根据误差获得异常分数。当重构帧或预测帧与原帧误差较大时，相应的异常分数也高，判断为异常发生的概率更大。但是这些误差的计算通常是基于整幅画面的，若异常发生的面积较小，则整体差异不大，获得的异常分数也偏低，容易产生漏检。

针对上述问题，本章提出了一种基于时空一致性和误差密度的视频异常检测方法，利用时空一致性减少正常事件的不可预测性和画面的急剧变化导致的误检，利用误差密度减少小面积异常导致的漏检。该方法在网络结构，损失函数和训练方式等方面与其他基于神经网络的视频异常检测算法相独立，可仅作用于异常分数的计算阶段，做到了即插即用，简单且高效地解决了上述问题。

3.2 基于误差密度的异常分数计算方法

不论是基于重构误差的视频异常检测算法，还是基于预测误差的视频异常检测算法，判断异常与否的关键步骤都是评估生成帧与原帧之间的差异程度，异常发生的可能性与差异程度成正比。因此，如何评估两张图片之间的差异，使其能够对小面积的差异产生响应，成了克服小面积异常漏检问题的关键。这一节首先介绍常用的图像误差评估方法，然后在此基础上介绍本文提出的一种能对误差密度进行表示的评估方法，用来计算异常的分数。

3.2.1 图像误差评估方法

均方误差（Mean Square Error，MSE）是一种常用的误差计算方法，它可以衡量真实视频帧和预测或重构帧之间的相似性，计算公式见式 3-1。

$$MSE(x,\hat{x}) = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|x(i,j) - x(\hat{i},j)\|^2 \tag{3-1}$$

其中 x 和 $\hat{x}$ 分别表示大小为 $m\times n$ 的真实帧和生成帧，像素的取值范围为 $[0,1]$，其中生成帧是由重构网络或预测网络生成的结果。

峰值信噪比（Peak Signal to Noise Ratio，PSNR）也是一种广泛使用的方法，是在图像重建、去噪等领域里重要的质量评价指标，其公式为3-2。

$$PSNR(x,\hat{x}) = 10\log_{10}\frac{1}{MSE(x,\hat{x})} \tag{3-2}$$

当两帧之间差异越大时，MSE也越大，而相应的PSNR越小，表示对应帧发生异常的概率越大。当PSNR越大时，真实视频帧和生成帧之间的差异越小，对应帧发生异常的概率则较小。在实际使用中，通常使用归一化函数对不同的测试视频进行归一化，获得第i帧的异常分数 $s(i)$，具体的计算方法见式3-3。

$$s(i) = 1 - \frac{PSNR(x_i,\hat{x}_i) - \min PSNR(x_t,\hat{x}_t)}{\max PSNR(x_t,\hat{x}_t) - \min PSNR(x_t,\hat{x}_t)} \tag{3-3}$$

$s(i)$ 越高，说明对应第i帧为异常的概率越大。在应用时工作人员通常会根据实际情况设置一个阈值，当异常分数大于该阈值时，将对应帧标注为异常帧，表明该帧视频中包含异常行为。当异常分数小于该阈值时，将对应帧标注为正常。阈值的设计决定了异常检测的敏感程度，这一选择通常由实际应用场景决定，视频异常检测算法只给出每一帧的异常分数。

这些计算方法都是基于整张图片的差异进行求和，因此对局部的差异不敏感。当小范围的异常发生时，虽然异常对应区域的差异比较大，但总体的差异和较小，获得的异常分数仍较小，容易产生漏报。

3.2.2 基于误差密度的异常分数计算方法

当小范围的异常发生时，虽然基于整张图片求和的差异衡量指标不能体现出其差异程度，但是在该异常发生的范围内，差异的密度通常很大。针对这一问题，Nguyen等人提出了基于块的异常分数计算方法[4]，这种方法虽然克服了小面积异常的漏检问题，但是涉及矩阵的求逆，计算比较复杂。在此基础上，我们提出了基于池化思想的误差密度估计方法，可以使用简单的计算步骤实现对误差密度的估计。

卷积神经网络（Convolutional Neural Network，CNN）是由LeCun等人[58]在1998年提出的模型。该网络在2012年的ImageNet竞赛中取得了第一名的成绩，并远高于当时其他算法的表现，自此得到了研究人员的关注，在计算机各领域尤其是视觉方向获得了广泛应用。

卷积神经网络的基本操作有卷积、激活和池化。其中池化操作可以将特定

位置的值用一定范围内的数值统计值来表示，能在保留特征的情况下降低数据的维度。常用的池化操作有最大池化和平均池化。最大池化可以提取局部最大值，保留主要的特征，而平均池化可以提取局部平均值，也是进行特征压缩的常用方法。在进行平均池化操作时，利用滑动窗口对参数进行平均，得到该区域的对应特征。

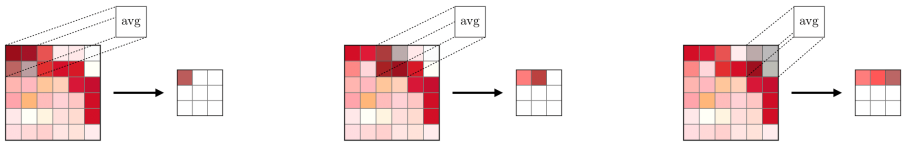


图 3-1: 平均池化过程示意图

图 3-1[①]展示的是平均池化操作。输入特征图的尺寸为 6×6，滑动窗口的大小为 2×2，步长为 2，补 0 个数为 0。在进行平均池化操作时，滑动窗口对相应的 4 个特征值求平均，得到一个值，作为该区域特征的表示，然后向右平移 2 步，对下一个区域继续求平均，如此往复。当特征图的前两行均被计算后，将滑动窗口平移至第三行第四行的最左侧，开始进行下一轮的计算，直至覆盖整个特征图。在进行平均池化操作以后，原特征图被压缩成 3×3 的特征图，尺寸变为原图的 $\frac{1}{4}$，其中每个特征对应着原图中 2×2 大小的区域的均值。

在 MSE 和 PSNR 的计算过程中，都是先对两张图片求差值，然后对差值求和，再进行其他的运算，这样得到的结果能有效的体现出两张图片的差异程度。但是这种求和的方式是基于整张图片的，不易检出小面积的异常。卷积神经网络的平均池化操作能够对区域的特征进行平均，求出对应区域的均值。因此，经过平均池化操作后得到的特征值能对一定区域的特征进行表示。我们可以对原监控视频图像和预测或者重构的图像进行求差，得到一个与原图尺寸一致的差值图 $DIFF(x_i, \hat{x}_i)$(式 3-4)。

$$DIFF(x_i, \hat{x}_i) = x_i - \hat{x}_i \tag{3-4}$$

然后，在差值图上做平均池化操作，滑动窗口的大小为 $k\times k$，步长为 k，补零个数为 0。经过平均池化操作，我们将获得一个压缩后的特征图 F（式 3-5）。这个图的面积比原图小，其中每个特征值 $F(m,n)$ 都对应着差值图

[①]图片来源：https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks

中大小为 $k \times k$ 的区域的平均值，这个值可以视为该区域的误差密度的估计。

$$F((x_i, \hat{x}_i), k) = Average\ Pooling(DIFF(x_i, \hat{x}_i), kernel\_size = k, strides = k, padding = 0) \tag{3-5}$$

当异常发生时，特征图 F 对应的值较大。若对 F 进行求和，则与 MSE 和 PSNR 思想一致，得到了基于全图的差异值的衡量。为了避免这类计算方法对小面积异常的漏检，我们在计算异常分数的时候，采用了求最大值的思想。在此情况下，只要有某一特征值 $F(m,n)$ 比较大，得到的异常分数就较高，这种异常分数的计算方式能够对小面积的异常进行响应，避免了上述问题。异常分数的计算过程如下，首先求出特征图 F 的最大值 f_i（式 3-6），然后对 f_i 进行归一化得到最终的 s_i（式 3-7），表示当前第 i 帧的异常分数。

$$f_i = MAX(F((x_i, \hat{x}_i), k)) \tag{3-6}$$

$$s_i = \frac{f_i - \min f_t}{\max f_t - \min f_t} \tag{3-7}$$

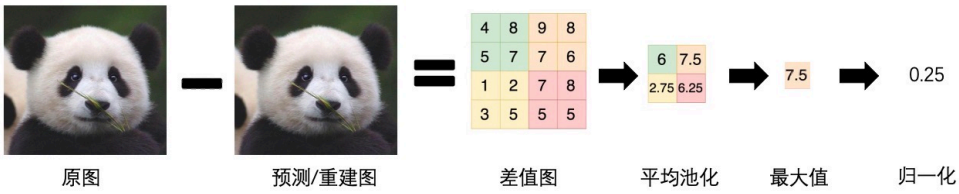


图 3-2: 基于误差密度的异常分数计算流程

图 3-2是异常分数计算的总体流程。由图可知，s_i 通过池化操作和一些简单的计算即可获得，计算比较简单，且能够有效的避免小面积异常的漏检问题。实验表明，在使用基于密度的异常分数计算方式时，获得的结果更鲁棒。尤其是当摄像机抖动的情况发生时，整张画面发生位移，预测帧或重构帧与原帧之间的差异之和会比较大，容易造成误检。若此时没有异常发生，此时整张差异图不会产生密度较高的点，基于误差密度的异常分数值仍较低，可避免误检的发生。

3.3 基于时间一致性的视频异常检测方法

在视频处理领域，时间一致性一直是一个值得关注的问题，不论是视频生成，风格转换还是视频着色，都需要使获得的结果在时间上具有一致性[59,60]。

直接在视频的每一帧上单独使用图像处理算法通常会产生在时间上不一致的结果。即虽然算法输出的每一帧都符合要求，但是在时间上不具有一致性，造成视频的观感不好，可能会存在模糊、抖动、闪烁等结果，影响最终的结果。因此，在视频处理领域，需要对视频的时间一致性进行额外的关注。

3.3.1 异常的时间一致性

视频中的异常也具有时间上的一致性。因此针对视频的异常检测算法需要对时间一致性问题进行关注。图 3-3为 CUHK Avenue 数据集[1] 中的奇怪行为的异常，画面中有一名男子在往空中扔书包，这种情况在人行道上是不多见的，因此被标记为异常行为。 可以看出，男子的行为在时间上具有一致性，其扔书

图 3-3: 异常事件的时间一致性

包的异常动作是连续的，相邻帧 x_i 和 x_{i+1} 之间具有很高的相似性，且 x_i 与 x_{i+1} 帧的异常面积大小也基本一致。

针对异常的一致性问题，当前基于深度学习的视频异常检测算法也做了一定的研究。最新的方法主要是在训练过程中对时间一致性进行约束，主流做法是对动作特征进行提取，使用诸如 Flownet，C3D[61] 等特征对相邻帧之间的动作进行约束，保证重构和预测的帧满足动作上的一致性。但是这种方法通常只考虑相邻两帧之间的时间一致性，且只对动作进行约束，不能保证静止物品的时间一致性。在此基础上，我们提出了一种基于时间一致性的异常分数计算方法，可以在非训练阶段保证异常的时间一致性，使视频的异常分数体现出这种一致性。

异常通常是持续一段时间的，很少存在只有一帧的异常。在相邻帧之间，异常在视野中的位置和面积也基本相似，因此基于差异图计算的异常分数也应具有时间上的一致性。在此情况下，我们可以充分利用异常事件的时间一致性来解决正常事件的不可预测性和画面急剧变化带来的误检问题。当不可预测的正常事件发生时，如车灯在第 i 帧突然打开。虽然预测帧 $\hat{x}_i$ 和实际帧 x_i 差异较大，该帧对应的异常分数会比较大，但是在车灯打开之前的第 $i-1$ 帧，预测帧 $\hat{x}_{i-1}$ 和实际帧 x_{i-1} 的差异不受影响。在车灯打开之后的 $i+1$ 帧，由于车灯

已经打开，预测器能很好的预测车灯打开的情况，预测帧 $\hat{x}_{i+1}$ 和实际帧 x_{i+1} 之间的差异亦不受影响。综上可知，基于差异图计算的异常分数在第 i 帧会突然变高，但第 s_{i-1} 和第 s_{i+1} 帧的异常分数不受影响。由此可得，当画面急剧变化时，异常分数也只会在变化的那一帧发生突变，但其他帧不受影响。因此，若使异常分数在时间上保持一致性，则能够避免正常事件的不可预测性问题和画面急剧变化问题带来的不好影响。

3.3.2 基于时间一致性的异常分数计算方法

由上一小节可知，视频中的异常通常具有时间一致性，基于差异图计算的异常分数也具有一致性。在现有的异常分数计算方法中，第 i 帧的异常分数仅与差异图 $DIFF(x_i, \hat{x}_i)$ 有关。考虑到异常的时间一致性，我们令第 i 帧的异常分数与前后 $2k$ 帧有关，即与 $DIFF(x_{i-k}, \hat{x}_{i-k})$,$DIFF(x_{i-k-1}, \hat{x}_{i-k-1})$,$\cdots$,$DIFF(x_{i-1}, \hat{x}_{i-1})$ 及 $DIFF(x_{i+1}, \hat{x}_{i+1})$,$\cdots$
,$DIFF(x_{i+k-1}, \hat{x}_{i+k-1})$,$DIFF(x_{i+k}, \hat{x}_{i+k})$ 有关。

高斯分布也称为正态分布，是数学、物理和工程等领域中最常见的分布之一，一维正态分布的公式见3-8。其中 μ 是分布的均值，σ^2 是分布的方差。当 $\mu = 0$ 时，一维正态曲线关于 y 轴对称，呈现两头高，中间低的形状。由视频的时间一致性可知，第 $i-1$ 和第 $i+1$ 帧与第 i 帧的相关性最强，第 $i-1$ 和第 $i+1$ 帧的相关性次之，随着与第 i 帧距离的增加，其他帧的相关性逐渐降低。因此，我们假设与第 i 帧最近的前后 $2k$ 帧对第 i 帧的影响满足高斯分布。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3-8}$$

记第 i 时刻生成帧与原帧的差异程度为 ano_score_i，ano_score_i 与差异图 $DIFF(x_i, \hat{x}_i)$ 有关，可以是第3.2.1小节中的基于 MSE 或 PSNR 计算得到的异常分数，可以是第3.2.2小节中基于误差密度的异常分数，也可以是通过其他方法计算得到的差异分数。ano_score_i 可以形式化的表示为：输入为原视频帧 x_i 与预测或重构帧 $\hat{x}_i$ 的差异图 $DIFF(x_i, \hat{x}_i)$，通过某种函数 $compute\_score$，获得对应的异常程度的分数，用来衡量第 i 帧发生异常的可能性（式3-9）。

$$ano\_score_i = compute\_score(DIFF(x_i, \hat{x}_i)) \tag{3-9}$$

第 i 帧最终输出的异常分数 $final\_ano\_score_i$ 为 ano_score_{i-k} 到 ano_score_{i-k}

的加权和。$w_{i,j}$ 表示在计算第 i 帧的最终异常分数时，第 j 帧对应的权重，由高斯分布获得。k 和 σ 是超参数。k 决定了第 i 帧参考帧的数目，k 越大，计算时参考的帧越多，获得的结果越平滑。σ 是高斯分布的标准差，决定了第 i 帧参考帧的权重，σ 越大，第 i 帧所占的权重越小，距离第 i 帧较远的帧所占的权重越大，获得的结果越平滑。得到 $w_{i,j}$ 后，将前面得到的 ano_score_{i-k}，ano_score_{i-k+1}，$\cdots$，ano_score_{i+k-1}，ano_score_{i-k} 带入式 3-11，获得最终的异常分数。

$$w_{i,j} = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(j-i)^2}{2\sigma^2}) \tag{3-10}$$

$$final\_ano\_score_i = \frac{\sum_{n=i-k}^{i+k} w_{n,m} \times ano\_score_n}{\sum_{n=i-k}^{i+k} w_{n,m}} \tag{3-11}$$

在计算异常分数时，需要对每段视频的前后 $k-1$ 帧进行特殊处理。当 $i < k$ 时，由于第 i 帧之前并不存在完整的 k 帧，在计算时只需要对前面 $i-1$ 帧的异常分值进行加权求和，计算公式见式 3-12。

$$final\_ano\_score_i = \frac{\sum_{n=1}^{i+k} w_{n,m} \times ano\_score_n}{\sum_{j=1}^{i+k} w_{n,m}} \tag{3-12}$$

当 $i > k$ 时，由于该帧之后并不存在完整的 k 帧，在计算时也要进行特殊的处理，计算公式见式 3-13。其中 n_{frame} 是测试视频的总帧长。

$$final\_ano\_score_i = \frac{\sum_{n=i-k}^{n_{frame}} w_{n,m} \times ano\_score_n}{\sum_{n=i-k}^{n_{frame}} w_{n,m}} \tag{3-13}$$

基于时间一致性的异常分数计算方法能够利用前后 k 帧的信息，对当前帧的异常情况进行判断，避免了基于单帧判断时易受噪声影响产生波动的情况。该方法可以视为对异常分数在时间维度上做高斯平滑操作，能够有效地去除噪声，避免单帧预测或重构时出错导致的误判，可以显著提高算法的鲁棒性。

3.4 实验设计与分析

3.4.1 实验设置

本节对上述两种异常分数计算方法的有效性进行了验证，并综合使用两种异常分数的计算方法，与两个不同的算法结合，分别在两种不同的数据集上与近年来的一些算法进行比较。

本实验首先分别对异常密度的有效性和时间一致性的有效性进行验证，然后将二者结合起来，验证其有效性。由第2.1小节中可知，Avenue 数据集中含有摄像头抖动的视频，是一个很难处理的挑战，而利用异常密度的分数计算方法能够有效的缓解该问题。因此，我们选择在 Avenue 数据集上对上述异常分数计算方法的有效性进行实验。在实验过程中，我们将异常分数计算方法和 FramePred 算法相结合。实验证明，分别使用方法能够提高算法的鲁棒性，使性能得到大幅提升。

最后，我们同时使用这两种异常分数计算方法，与现有的视频异常检测方法相结合，并与当前比较好的算法进行比较：一是将计算方法与 FramePred 算法进行结合，在 Avenue 数据集上进行实验；二是将计算方法与 MNAD 算法相结合，在 Avenue 和 Ped2 数据集上进行实验。实验首先对训练数据和测试数据进行预处理，将图像大小调整到 256×256，然后将像素值归一化到 $[0, 1]$ 范围内，获得 X_{train} 和 X_{test}。训练流程如图3-4所示，这是一种端到端的训练方法，只需要将测试数据输入到网络中去，然后对网络进行训练获得相应的模型。具体参数和流程与 FramePred 算法和 MNAD 算法论文一致。在测试阶段 (图3-5)，我们将 X_{test} 输入到模型中去，获得 $\hat{X}_{test}$，然后计算他们的差分图 DIFF，即 $DIFF(x_i, \hat{x}_i)$，其中 $i = 1...n_{train}$，之后利用图3-2中的计算过程获得基于误差密度的异常分数 ano_score_i，最后利用公式3-11获得最终基于误差密度和时间一致性的异常分数 $final\_ano\_score_i$。

图 3-4: 实验的训练阶段

算法使用了 Pytorch 框架，运行的 GPU 为 NVIDIA 1080-Ti。

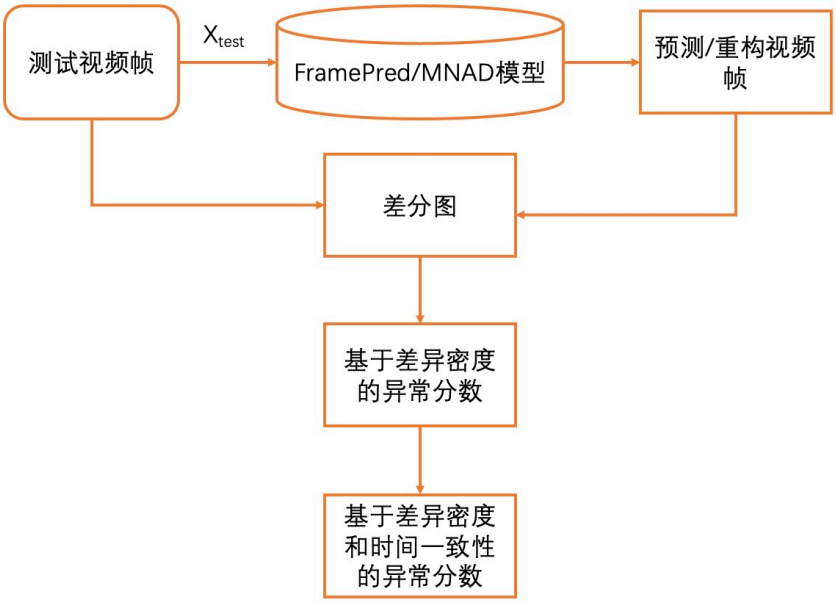


图 3-5: 实验的测试阶段

3.4.2 异常密度的有效性验证

异常密度是针对小面积异常不易检出问题提出的解决方案。除此之外，当摄像头抖动等事件造成大面积异常，但局部异常密度不高时，使用基于异常密度的视频异常检测算法能够降低误报率。Avenue 数据集中含有部分抖动的视频，因此，我们将基于异常密度的异常分数计算方法与 FramePred 算法相结合，在 Avenue 数据集上验证异常密度的有效性。实验时的训练参数同论文[5]，共训练 10000 轮，异常分数的计算方法为基于误差密度的异常分数计算方法。

平均池化过程中核大小 k 的选择非常重要，当 k 的值过大时，不能对局部的异常密度进行体现，计算过程接近于基于全局求差异和的方式；当 k 的取值过小时，又退化成基于像素层级的差异计算，会造成非常大的抖动，极易被单个噪声点的取值所影响。在实验过程中，我们将 k 设置为 16，既不会因为过小导致性能不稳定，又不会因为过大影响了对局部信息的敏感性。

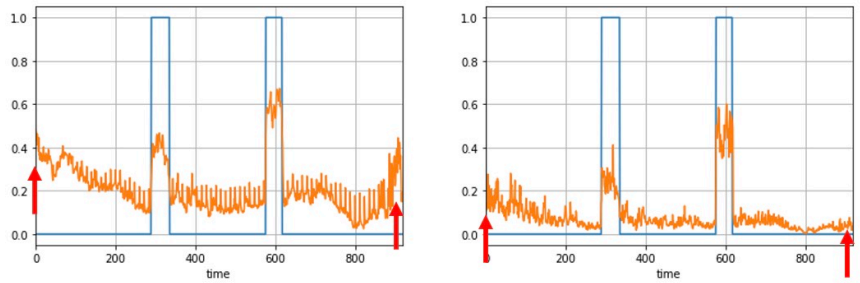
图 3-6展示了异常密度与异常的相关性，每张图片均为真实的视频场景与差异图的拼接。左图拍摄时镜头发生了抖动，此时有一些人在行走，无异常行为发生。可以看出，当镜头抖动发生时，虽然整张差异图全局差异比较大，有很多有差异的地方，但是总体而言误差密度都比较小。而右图是真正发生异常时的情况，红色方框中的人在奔跑。此时右侧的差异图中有误差密度比较大的

部分。由此可知，当异常发生时，不仅基于整张差异图的全局差异较大，局部异常密度也会比较大。

镜头抖动时无密度特别大的地方　　异常发生时有密度特别大的地方

图 3-6: 异常密度与异常的相关性

在未使用基于误差密度，而是采用 FramePred 中提出的异常分数计算方法时，算法在 Avenue 测试集中第 3 个测试视频上的异常分数结果如图 3-7中左图所示。图中横轴是视频的帧数，黄色的每一点对应着归一化到 [0, 1] 范围内的



测试视频3的异常分数　　测试视频3基于差异密度的异常分数

图 3-7: 在镜头抖动的情况下使用基于误差密度的异常分数计算方法

异常分数，异常分数越高，该帧发生异常的概率越大。蓝色的值是真实的异常标签，当标签为 1 时表示该帧有异常发生。在视频的开头和结尾处摄像机镜头发生抖动，分别被红色箭头指出。可以看出，在抖动发生时，虽然无异常发生，但异常分数有明显的提升。

在使用基于误差密度的异常分数计算方法之后，获得的异常分数如图 3-7中右图所示。我们可以看到在箭头所指之处，异常分数明显降低，而真正异常发生时的异常分数基本保持不变，而其他时间的异常分数基本保持原有趋势。由此可以看出，基于异常密度的异常分数计算方式能够明显降低由于镜头抖动带来误检的可能性。

图 3-8展示了异常的面积较小时的情况。该图为 Avenue 数据数据集中第 1 个测试视频的第 109 帧，此时有人在快速奔跑。由于奔跑的人距离摄像头较

远，异常在整幅画面中所占比例相对较小。此时利用 FramePred 算法直接对该画面所属的测试视频进行检测，基于 PSNR 计算得到图 3-9中左图所示的异常分数，其中绿色的线表示整个视频异常分数的均值。图中红色箭头指出的地方是异常发生的时刻，可以看出此时异常分数比均值低，且与该异常结束后的异常分数没有明显区分。在这种情况下，算法不能很好的分辨出异常的发生，容易造成漏检。使用基于误差密度的异常分数计算方法获得的输出如图 3-9中右图所示。可以看出，图中红色箭头所指的地方已经高于异常分数的均值，且与异常发生时刻前后帧的异常分数有较大的区分度，能够更好的检测出异常。

图 3-8: 异常面积相对较小的情况

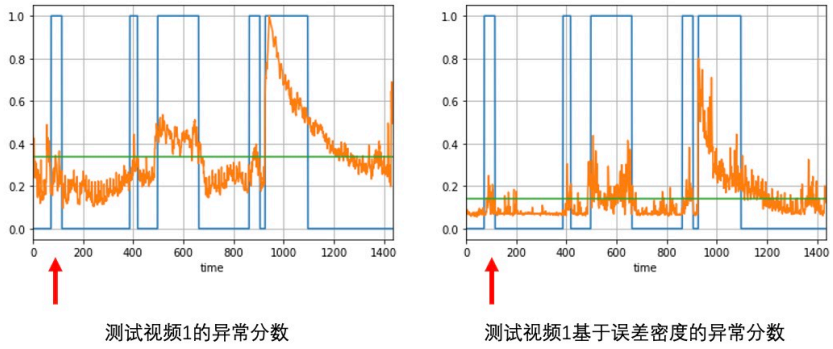


图 3-9: 在小面积异常的情况下使用基于误差密度的异常分数计算方法

3.4.3 时间一致性的有效性验证

利用时间一致性能够缓解由突发正常事件带来的误检问题，通过对前后帧误差的观察来判断当前帧的异常程度。在进行时间一致性的异常分数计算时，

我们令 $k = 11$，即某一帧的异常分数与其前后各 11 帧相关。式3-10中的 σ 设为 9。实验的数据集仍是 Avenue，算法采用 FramePred。

当普通异常分数计算方法出现短时间误报，即第2.2.1小节中的的 FP 时，当前帧的差异图对应的差异值较大。此时若前后 k 帧中有一些是 TN，对应的异常分数比较低。基于时间一致性的异常分数是基于前后 k 帧的异常值决定的，可以使当前帧的异常分数被平均后变低，在一定程度上缓解误报问题。当普通异常分数计算方法出现短时间的漏报，即 FN 时，若前后帧中有 TP，而这些帧异常分数较高，使用基于时间一致性的异常分数计算方法获得的异常分数被拉高，能够缓解漏报问题。

图3-10中左图展示的是在测试视频 4 上未使用基于时间一致性的异常分数。红色箭头展示的地方为误报。 右图展示的是使用基于时间一致性后得到

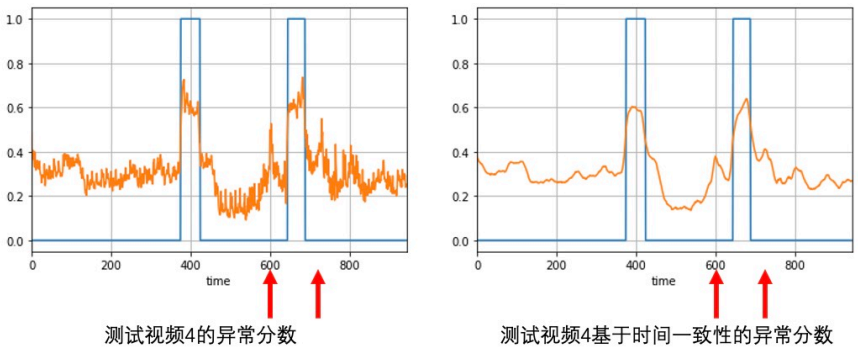


图 3-10: 使用基于时间一致性的异常分数计算方法

的异常分数。可以看出，红色箭头指向的地方有明显的降低，说明基于时间一致性的异常分数能够有效地处理这种情况。另外我们可以看出，基于时间一致性的异常分数计算方法等效于对异常分数做了时间维度上的平滑。由此不难推出，当出现噪声时，也能通过这种方式进行处理，以降低噪声带来的不良影响，有效地提高了视频异常检测系统的稳定性。

3.4.4 误差密度和时间一致性相结合的方法

误差密度和时间一致性在异常分数的计算中既可以单独使用，用来解决其单独针对的问题，也可以结合在一起使用，综合提高算法的检测水平。由于误差密度是基于最大值函数计算的，因此可能会被噪声影响。当图中某一区域由于噪声差异较大时，虽然没有异常发生，也会产生较大的异常分数，而基于差

异和的异常分数计算方法不会受到局部噪声的影响，相对比较稳定。当时间一致性被引入时，可以降低误差密度引入的不稳定性，将基于密度误差和时间一致性的异常分数计算方法结合在一起可以获得更好的结果。

在计算过程中，我们将两种方法串行使用，先使用基于误差密度的异常分数计算方法，再使用基于时间一致性的异常分数计算方法，获得最终的异常分数。算法的输入为 X_{test}，首先利用图 3-2中的计算过程获得基于误差密度的异常分数 ano_score_i。然后将该分数带入公式 3-11、 3-12和 3-13获得最终的异常分数 $final\_ano\_score_i$。在计算基于误差密度的异常分数时，我们令核的大小为 16。在计算基于时间一致性的异常分数时，我们令 $k = 11$，$\sigma = 9$。

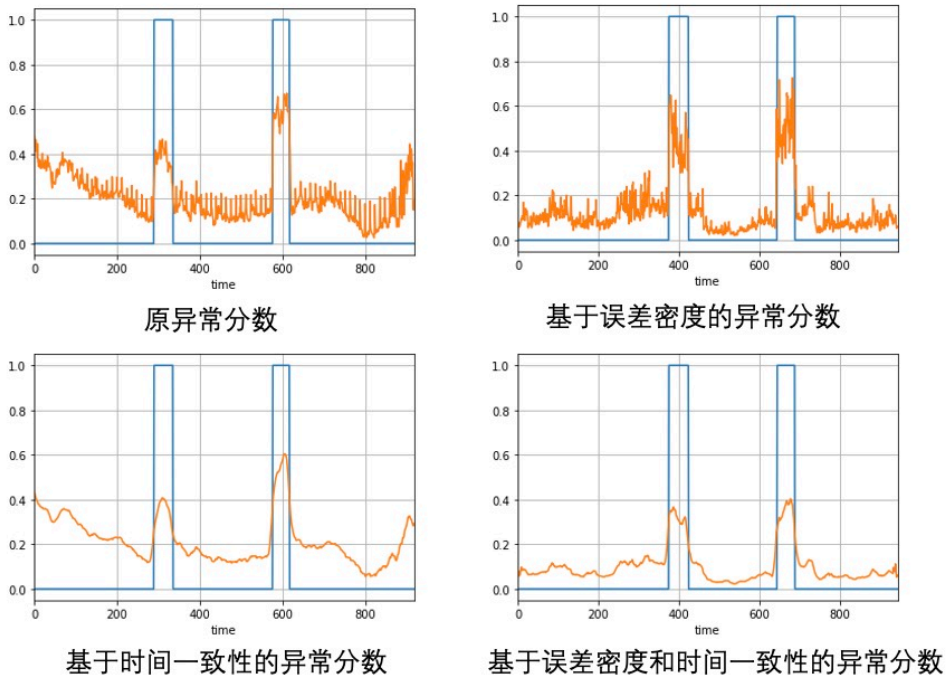


图 3-11: 误差密度和时间一致性相结合的异常分数

图 3-11展示了误差密度和时间一致性相结合的方法在 Avenue 数据集第 3 个测试视频上的效果。左上角的图像是 FramePred 算法输出的异常分数。右上角的图像是 FramePred 算法与基于误差密度的异常分数计算方法相结合产生的输出。左下角的图像是基于时间一致性的异常分数。右下角的图像是基于异常密度和时间一致性的异常分数。由图像可以看出，单独使用基于差异密度的异常分数计算方法时，能够抑制正常情况下的异常分数过高的情况，使得异常和正常行为在分数上更有区分度，但是获得的异常分数仍像原始异常分数一样，

有较大的抖动。单独使用基于时间一致性的异常分数时，分数能够得到有效的平滑，减少因为异常分数抖动造成的误报，但是异常和正常行为在区分度上并没有使用基于误差密度的异常分数计算方法好。同时使用这两种方法可以结合二者的优点，获得一个更鲁棒更准确的结果。

3.4.5 与其他方法的比较

基于差异密度和时间一致性的异常分数计算方法只对异常分数计算阶段进行了改进，因此是一个可以和其他算法相结合，并做到即插即用的算法，为了验证这两种计算方法的有效性，我们分别将他们与 Liu 等人提出的基于帧预测的算法（FramePred）和 Hyunjong 提出的基于记忆模块的算法（MNAD) 结合在一起，对其有效性进行验证。实验数据集分别为 Avenue 数据集和 UCSD Ped2 数据集，这些数据集已经在第 2.1节进行了详细介绍，此处就不再赘述。实验的评估指标是基于帧的 AUC，AUC 值越高，算法的效果越好。

我们在 Avenue 数据集上基于分别基于 FramePred[5] 和 MNAD[7] 进行异常分数的计算，并将得到的结果与方法 ConvAE[14]，TSC[62]，StackRNN[62]，MemAE[63] 和 MNAD[7] 进行比较。基于 FramePred 的算法参数设置同 FramePred 论文，训练时使用 Adam 优化器[64]，其中生成器的学习率为 0.0002，判别器的学习率为 0.00002，$\beta_1 = 0.9$，$\beta_2 = 0.999$，batch 大小为 4，共训练 10000 个 epoch。基于 MNAD 的算法参数设置同 MNAD 论文，训练时使用基于预测的方法，优化器为 Adam，其中学习率为 0.0002，$\beta_1 = 0.9$，$\beta_2 = 0.999$，batch 大小为 4，共训练 60 个 epoch。最终获得的 AUC 如表 3-1所示。

可以看出使用了基于误差密度和时间一致性的异常分数计算方法后，FramePred 的 AUC 比原方法有明显提升，与其他最新算法相比也有一定的竞争力。MNAD 的 AUC 也比原方法有明显提升，且与其他算法相比获得了最好的结果。由此可以看出，基于误差密度和时间一致性的异常分数计算方法在与 PramePred 和 MNAD 结合时，在 Avenue 数据集上能够取得性能的提升。

Avenue 数据集上包含镜头抖动的场景，为了验证该计算方法在不含镜头抖动的场景下依然有效，我们在 Ped2 数据集上基于 MNAD[7] 进行异常分数的计算，并将获得的结果与方法 ConvAE[14]，TSC[62]，StackRNN[62]，MemAE[63] 和 FramePred 进行比较。实验的参数设置同 MNAD 论文，训练时使用基于预测的方法，优化器为 Adam，其中学习率为 0.0002，$\beta_1 = 0.9$，$\beta_2 = 0.999$，batch 大小为 4，共训练 60 个 epoch。获得的 AUC 如表 3-2所示。将 MNAD 与基于误

表 3-1: 异常分数计算方法在 Avenue 数据集上的 AUC 结果

方法	AUC
ConvAE[14]	80.0
TSC[62]	80.6
StackRNN[62]	81.7
FramePred[5]	85.1
MemAE[63]	83.3
MNAD[7]	88.5
FramePred+ours	86.4
MNAD+ours	89.9

表 3-2: 异常分数计算方法在 Ped2 数据集上的 AUC 结果

方法	AUC
ConvAE[14]	85.0
TSC[62]	91.0
StackRNN[62]	92.2
FramePred[5]	95.4
MemAE[63]	91.7
MNAD[7]	97.0
MNAD+ours	98.5

差密度和时间一致性的异常分数计算方法结合在一起，可以显著的提高算法在 Ped2 上的检测水平，并在这一系列算法中获得了最优的结果。

由此，我们可以看出: 基于误差密度和时间一致性的异常分数计算方法具有即插即用性，非常方便使用，该方法可以和现有的视频异常检测算法结合起来，在异常分数计算阶段发挥作用；基于误差密度和时间一致性的异常分数计算方法能够提高现有算法在视频异常检测算法的性能，获得更准确的异常程度判断结果；基于误差密度和时间一致性的异常分数计算方法在与 MNAD 结合时，同当下顶尖算法相比能够取得最好的结果。

3.5 本章小结

本章提出了一种基于误差密度和时间一致性的异常分数计算方法，该方法可用于视频异常检测的异常分数计算阶段，与神经网络的训练和测试阶段相独立，可以做到即插即用。当前基于深度学习的视频异常检测算法主要是利用预测帧与实际视频帧之间的误差估计异常发生的可能性。基于误差密度的异常分数计算方法参考了卷积神经网络的平均池化过程，能够对原帧和预测帧或重构帧之间的差异图进行压缩，对局部的误差密度进行描述。该方法与基于全图差异求和的异常分数计算方式有明显区分，是对密度的最大值进行衡量，可以对小面积的较大误差密度进行响应，够有效地避免小面积异常漏检问题。基于时间一致性的异常分数计算方法在计算分数时考虑前后帧的时间一致性，通过前后帧来估算当前帧的异常程度，能够缓解由于突发正常事件带来的不良影响，降低噪声点带来误判的可能性。在上述研究的基础上，本章设计了一系列实验，证明了基于误差密度和时间一致性的异常分数计算方法的有效性。

第四章　基于内容损失和误差密度损失的损失函数设计

在第三章中，主要是对异常分数的计算方法进行了改进，使得获得的异常分数更鲁棒，准确性更高。这一做法是即插即用的，可以和现有的其他模型进行搭配使用，获得更好的效果。除此之外，我们还对训练过程中的损失函数设计进行研究，对现有的损失函数进行调研，并作出了补充，提出了内容损失和误差密度损失，对输出帧与实际帧之间的内容和误差密度进行约束，在视频异常检测任务中实现性能的提升。

4.1　常用损失的局限性

在基于神经网络的视频异常检测任务中，在训练阶段会利用损失函数对预测帧或重构帧进行约束，使得预测、重构帧与原帧尽量接近。这样在预测阶段，正常情况下的差异图 $DIFF(x_i, \hat{x}_i)$ 会更小，获得的异常分数也变小。常用的损失函数有以下几类：

首先是基于图像像素的损失函数：强度损失函数（Intensity Loss) 和梯度损失函数（gradient Loss），它们是对原帧和预测、重构帧之间外观表示的直接约束。强度损失函数的定义见式 4-1。它直接对像素差进行计算，是两张图像之间的 L2 距离。强度损失函数的约束可以使得两张图像的各点像素比较接近。

$$L_{int}(x_i, \hat{x}_i) = \|x_i - \hat{x}_i\|_2^2 \tag{4-1}$$

梯度损失函数是对图像的梯度进行约束，可以使生成的图像边缘形状更接近，更加锐化。梯度损失函数的定义见式 4-2，其中 m, n 是图像空间上的索引，$x_{i,m,n}$ 表示原视频图像第 i 帧第 m 行第 n 列位置的像素，$\hat{x}_{i,m,n}$ 表示重构或预

测图像第 i 帧第 m 行第 n 列位置的像素：

$$L_{gd}(x_i, \hat{x}_i) = \sum_{m,n} \| |\hat{x}_{i,m,n} - \hat{x}_{i,m-1,n}| - |x_{i,m,n} - x_{i,m-1,n}| \|_1 + \| |\hat{x}_{i,m,n} - \hat{x}_{i,m,n-1}| - |x_{i,m,n} - x_{i,m,n-1}| \|_1 \tag{4-2}$$

除此之外，还有基于运动信息的约束，可以通过这一约束实现运动上的一致性，其中光流是一种常用的约束。我们可以通过 FlowNet，FlowNet2[65]，TVNet[66] 等网络对第 i 帧和第 $i+1$ 帧的光流进行计算，获得相应的光流 $OpticalFlow(x_i, x_{i+1})$。光流损失的定义见式 4-3。由公式可知，光流损失先计算预测或重构的第 i 帧与实际第 $i+1$ 帧之间的光流，再计算实际第 i 帧与第 $i+1$ 帧之间的光流，使其越接近越好。这种约束能够对视频中运动的物体进行描述，保持物体运动的前后一致性。

$$L_{op} = \|OpticalFlow(\hat{x}_i, x_{i+1}) - OpticalFlow(x_i, x_{i+1})\| \tag{4-3}$$

上述损失虽然能够很好的对视频帧的外观和运动进行约束，但是针对外观的约束主要是一整个图像上的求和，不能很好地对局部差异进行响应。而且不论是针对外观的约束，还是针对运动的约束，都没有对视频图像的具体内容进行理解，没有在内容层面进行相应的约束。当摄像头发生抖动时，网络的外观损失会很大，运动损失也会很大，但是由于实际上并未发生异常，内容损失相应较小，不会产生误报。除此之外，基于内容的损失约束能够对视频内容进行更好的理解，保证生成的预测或重建帧与原视频帧具有较为相似的内容。

4.2 内容损失

4.2.1 内容理解和迁移学习

随着深度神经网络的迅速发展，其在图像领域的应用越来越广泛，在图像语义分割[67]，图像分类[68]，目标检测[69]，图像生成[70] 等领域均能取得不错的效果。Zhou 等人[9] 通过对神经网络的可解释性研究，提出了 CAM(Class Activation Mapping)，发现卷积神经网络不仅具有很强的图像处理和分类能力，还能对图像中的关键部分进行定位。图 4-1展示了部分卷积神经网络的对应的 CAM，根据热力图可以看出，这些深度特征能够对图像内容进行理解和定位。因此，我们可以在视频异常检测的过程中，通过对视频帧进行深度特征的提

取，获得对监控画面内容的理解，提高检测的水平。Zhou 等人还指出，深度特征的理解定位能力可以被迁移到其他数据集上，也可以用于通用的分类、定位等任务。

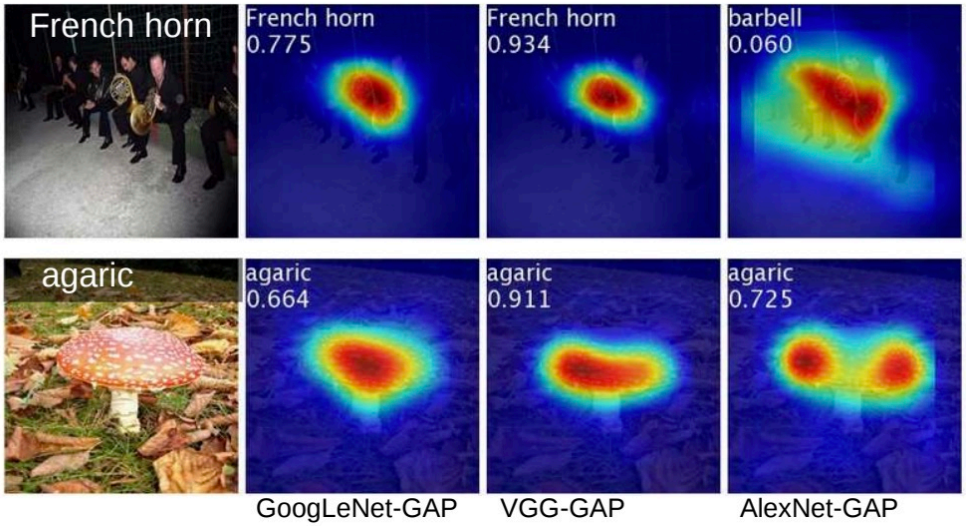


图 4-1: 部分卷积神经网络的深度特征[9]

由第1.1.3小节可知，在视频异常检测领域大规模的视频训练数据是非常缺乏的。因此，使用神经网络对视频异常检测相关数据集进行训练，获得其内容信息的描述非常困难。迁移学习是指两个不同领域的知识的迁移过程，可以利用源领域 D_s 的数据集训练模型，然后将学到的知识迁移到目标领域 D_t 上的学习任务。近年来，迁移学习被广泛应用于深度学习领域，尤其是图像处理领域[71]。在此情况下，我们可以采用迁移学习的方式，利用计算机视觉尤其是图像领域已经有的大规模高质量数据集，学习到一些可以泛化的知识。

迁移学习可以分为两个类型：归纳迁移学习（Inductive Transfer Learning）和推导迁移学习（Transductive Transfer Learning）。其中归纳迁移学习中基于特征的方式是将预训练模型提取的特征输入到目标任务的学习模型中[72]。这一特征可以是预训练模型的输出，也可以是中间隐藏层的输出。深层神经网络不同层的可迁移性有所不同[73]。一般来说，网络的底层学到的是通用的底层特征，如纹理，线条等，中层或者高层学到的是抽象的高级语义特征，而最后几层通常与特定任务相关。因此，在进行基于特征的归纳迁移学习时，可根据目标任务的特点来选择不同层的特征，以达到比较好的效果。

4.2.2 基于迁移学习的内容损失

ImageNet[74] 拥有超过 1400 万个标注图像，是计算机图像领域最著名的数据集之一。基于 ImageNet 的大规模视觉识别挑战赛 (ImageNet Large Scale Visual Recognition Challenge，ILSVRC) 通过目标识别和图像分类任务对算法进行评估，成为研究者对不同算法之间的性能差异进行比较的基准。因此，我们选择在 ImageNet 数据集上对深度神经网络进行预训练。

在计算机图像领域，LeNet[58]、AlexNet[75]、VGG[76]、Inception[77]、ResNet[10] 等神经网络均支持深度迁移学习。近些年，又涌现了许多新的神经网络，如 Wide ResNet[78]，MobileNet[79]，MNASNet[80]，和 Efficientnet[81]。由于不同的深度神经网络结构设计和损失选择略有不同，即便在同样的训练集上进行训练，获得的深度特征也有所不同。

ResNet 使用残差网络解决了增加深度带来的退化问题，使得网络更易于优化，且能够通过增加网络深度来提高准确率。Kolesnikov 等人[82] 提出了大迁移的思想（Big Transfer，Bit)，利用 ResNet 作为主干模型，训练出 BiT-S、BiT-M 和 BiT-L 三种规模的预训练模型，这些模型在二十多个数据集上都取得了很好的效果。其中，BiT-S 和 BiT-M 均是在不同规模的 ImageNet 进行预训练的。这说明在 ImageNet 上对 ResNet 进行训练能够取得很好的迁移效果。因此，我们也采用在 ImageNet 上预训练过的 ResNet 模型进行内容特征的提取。

为了在保证精度的情况下获得较快的速度，我们选择了相对轻量的 ResNet34 作为基础的内容特征提取网络。ResNet 使用残差学习来解决神经网络训练过程中的退化问题。图 4-2为 ResNet 的基本结构：残差学习单元。该单元在网络中增加了短路连接（shortcut connection)，可以对残差进行学习。

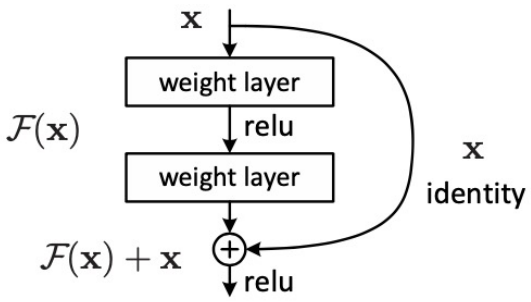


图 4-2: 残差学习单元[10]

ResNet34 的具体网络结构如图 4-3中右图所示，可以看到，它在左图的基

础上做了一些改进，在每两层之间加上短路连接，形成残差学习单元。

我们取最后一个全连接层的输出作为最终的深度特征，用来对图像的内容进行描述，得到一个 1000 维的特征。将实际视频帧 x_i 和预测或重构帧 $\hat{x}_i$ 分别输入在 ImageNet 上预训练好的 ResNet34 网络，获得对应的内容描述 $content_i$ 和 $\hat{content}_i$:

$$content_i = ResNet34(x_i) \tag{4-4}$$

$$\hat{content}_i = ResNet34(\hat{x}_i) \tag{4-5}$$

在训练的过程中，我们引入内容损失（式4-6），对输出帧 $\hat{x}_i$ 与实际帧 x_i 进行内容上的约束。当输出帧与实际帧之间的内容差异较大时，内容损失 $L_{content}$ 也相应变大，导致网络整体损失增大。通过引入内容损失对视频的内容信息进行约束，可以提高预测帧或者重构帧的质量，进而提高视频异常检测算法的性能。

$$L_{content}(x_i, \hat{x}_i) = ||content_i - \hat{content}_i|| \tag{4-6}$$

4.3 误差密度损失

由第4.1节可知，当前主流算法中几种常见的损失函数是基于整张图的差异和计算的。当图片在较小面积内产生差异时，损失函数往往不能对其产生响应。在使用这些损失对自编码器进行训练时，若整张图的重构效果不错，但有一些细节处理的不够好时，整体的损失函数仍比较小。在这种情况下，训练出来的网络在小面积细节处理方面将略有欠缺。由第3.2.2节可知，基于误差密度的异常分数计算方法可以有效检测出误差密度大的区域，当细节差异较大时，也会产生比较大的差异值。因此，在训练过程中，可以引入误差密度对生成帧和实际帧进行约束，使生成帧与实际帧之间的每一个小区域都较为相近。

我们引入误差密度损失对图片的细节进行约束，计算流程与基于误差密度的计算方法相类似，但是不需要对最终的分数进行归一化。实际视频帧 x_i 和预测或重构帧 $\hat{x}_i$ 对应的误差密度损失计算步骤如下:

1. 对两帧求差异图，获得 $DIFF(x_i, \hat{x}_i)$
2. 利用公式3-5求特征图 $F((x_i, \hat{x}_i), k)$
3. 用特征值的最大值作为损失函数:

$$L_{diffDensity} = MAX(F((x_i, \hat{x}_i), k)) \tag{4-7}$$

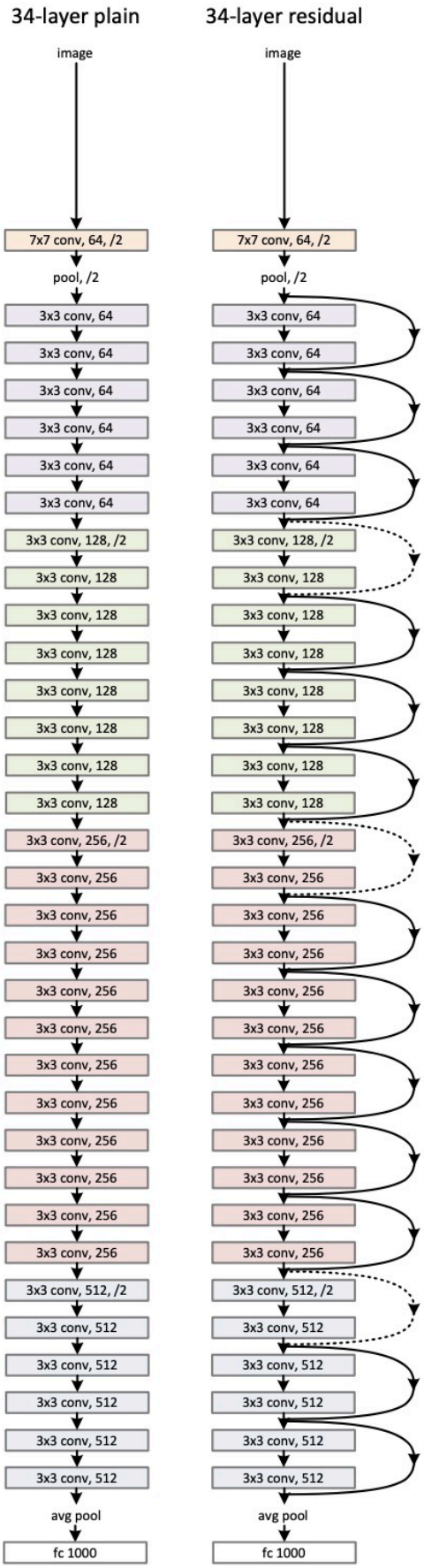


图 4-3: ResNet34 网络结构[10]

在训练的过程中使用误差密度损失可以对生成图像的细节进行约束，得到小范围内更接近真实图像的输出。

4.4 实验设计与分析

4.4.1 实验设计

本实验参考了 FramePred[5] 的的思想，是一种基于预测误差的视频异常检测方法。算法利用 U-net 网络[47] 进行帧的预测，输入是从第 i 帧开始的 p 个连续视频帧 $x_i x_{i+1} ... x_{i+p-1}$，输出是 $\hat{x}_{i+p}$。传统的图像生成网络通常由两个模块组成，一个是解码器，用来将提取图像的特征；另一个是生成器，利用特征进行上采样，获得图像。然而这种网络容易出现梯度消失现象，为了解决这一问题，U-net 在解码器和生成器之间加入短路链接，抑制梯度消失。图 4-4展示了 U-net 网络的基本结构。其中红色的箭头表示最大池化，蓝色的箭头表示卷积操作，黑色的箭头表示反卷积操作，绿色的箭头表示拼接操作。

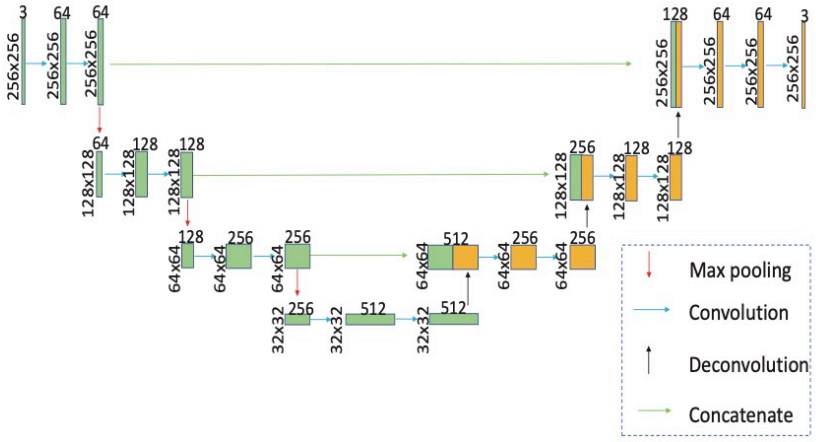


图 4-4: U-Net 的网络结构[5]

实验采用了基于对抗生成网络的训练方法，利用预测模块作为生成器，输出预测帧 $\hat{x}_{i+p}$，利用 PatchGan[70] 中提出的判别器，判断输入的是预测帧 $\hat{x}_{i+p}$ 还是真实的下一帧 x_{i+p}。PatchGan 使用了基于块的思想，它完全由卷积层构成。图 4-5展示了它与传统网络的不同。左图是传统网络使用的判别器，一般是对整个图片求最终的分数，而右图是 patchGAN 中使用的判别器，它对图片的各个部分给出判断，然后对这些判断求平均，获得整张图片的分数。它将输入的

图片转换为 64 个 128×128 的特征图，然后对其进行压缩，获得 256 个 8×8 的特征图，接着对这 256 个 patch 分别判断真假，取平均值作为最终的结果。在训练的过程中，生成器的目标是生成更好的预测帧，使判别器无法判断输入的是预测帧还是真实的下一帧。而判别器的目标是更好的区分这两种帧。

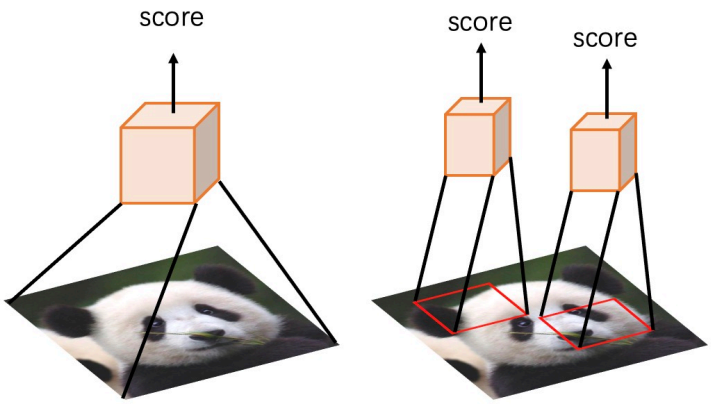


图 4-5: PatchGAN 的思想

生成网络的损失函数包含两个部分：一是对预测帧的约束，有强度损失 L_{int}，梯度损失 L_{gd}，光流损失 L_{op}，内容损失 $L_{content}$ 和误差密度损失 $L_{diffDensity}$；二是对抗损失，是对抗训练过程中的损失，目的是使预测网络获得的预测帧能够更好的骗过判别器。其中强度损失用来约束预测帧和真实帧像素级别的差异，梯度损失用来对图像的边缘进行约束，光流损失用来对运动的情况进行约束，内容损失用来对图像的内容进行约束，误差密度损失用来对局部误差密度进行约束。通过这五种不同的损失函数可以保证预测帧与实际帧在外观，边缘，运动，内容和局部都具有相似性。生成网络的总损失函数见公式 4-8，其中 λ_{int}，λ_{gd}，λ_{op}，λ_{adv}，$\lambda_{content}$ 和 $\lambda_{diffDensity}$ 为这些损失对应的权重。

$$\begin{aligned} L_g =& \lambda_{int} L_{int}(x_{i+p}, \hat{x}_{i+p}) \\ &+ \lambda_{gd} L_{gd}(x_{i+p}, \hat{x}_{i+p}) \\ &+ \lambda_{op} L_{op}(x_{i+p}, \hat{x}_{i+p}) \\ &+ \lambda_{adv} L^{G}_{adv}(\hat{x}_{i+p}) \\ &+ \lambda_{content} L_{content}(x_{i+p}, \hat{x}_{i+p}) \\ &+ \lambda_{diffDensity} L_{diffDensity}(x_{i+p}, \hat{x}_{i+p}) \end{aligned} \tag{4-8}$$

判别网络的损失函数如下：

$$L_d = L_{adv}^{D}(x_{i+p}, \hat{x}_{i+p}) \tag{4-9}$$

在测试阶段，算法利用训练得到的 U-net 网络对帧 $x_i x_{i+1}...x_{i+p-1}$ 进行预测，获得 $\hat{x}_{i+p}$。之后对 x_{i+p} 和 $\hat{x}_{i+p}$ 求差，得到 $DIFF(x_{i+p}, \hat{x}_{i+p})$，在此基础上获得基于误差密度的异常分数 ano_score_{i+p}，最后利用公式3-11求基于误差密度和时间一致性的异常分数 $final\_ano\_score_{i+p}$。

4.4.2 定量分析

在实验的过程中，我们首先将所有的视频帧调整为大小 256×256 的图像，然后将其归一化到 $[0, 1]$ 范围内。训练过程中以连续的 5 帧作为训练数据，将前 4 帧输入到网络中，输出预测的第 5 帧。生成器的学习率为 0.0002，判别器的学习率为 0.00002，$\beta_1 = 0.9$，$\beta_2 = 0.999$，batch 大小为 4，共训练 80000 个 epoch。输入序列的长度 $p = 5$。并将损失函数的参数设置为：$\lambda_{int} = 2$，$\lambda_{gd} = 2$，$\lambda_{op} = 2$，$\lambda_{adv} = 0.05$，$\lambda_{content} = 2$，$\lambda_{diffDensity} = 2$。在计算视频的异常分数时，我们采用了基于误差密度和时间一致性的异常分数计算方法。并设置平均池化的核大小为 16，时间一致性的 $k = 11$，$\theta = 9$。实验的评价指标为 AUC，数据集为 Avenue，算法基于 Pytorch 框架，并在 NVIDIA 1080-Ti GPU 上运行。

我们将得到的结果与方法 ConvAE[14]，TSC[62]，StackRNN[62]，FramePred[5]，MemAE[63]，MNAD[7] 进行比较，得到表 4-1。其中，ConvAE，TSC，StackRNN 和 MemAE 是基于重构误差的视频异常检测算法，FramePred 是基于预测误差的视频异常检测算法。MNAD 既可以利用重构误差进行异常检测，也可以利用预测误差进行异常检测，但是实验表明，在 Avenue 数据集上，使用预测误差检测得到的结果更好，因此此处展示的是基于预测误差的检测结果。

由表可知，我们的方法在 Avenue 数据集上能够取得 88.3 的 AUC，在这些算法中排名第二，仅次于 MNAD。尽管本文提出的算法实验结果比 MNAD 稍差，但 MNAD 需要额外维护一个记忆模块。该模块需要对获得的隐向量进行记录，并在训练过程中不断更新其中的内容。记忆模块需要占用额外的空间来存储隐向量，同时要耗费计算资源对记忆进行管理。而我们的算法不需要对记忆模块进行维护，也节省了空间上的开销。

表 4-1: 算法在 Avenue 数据集上的 AUC 结果

方法	AUC
ConvAE[14]	80.0
TSC[62]	80.6
StackRNN[62]	81.7
FramePred[5]	85.1
MemAE[63]	83.3
MNAD[7]	88.5
Ours	88.3

与其他算法相比较，我们算法的表现则提升了 2.2 至 8.3，证明了该算法在视频异常检测问题上的有效性。我们的算法在训练阶段引入两种额外的损失，对训练结果进行约束。而在测试阶段，算法并未对网络进行修改，仅需使用 U-Net 网络对视频帧进行预测即可，且基于差异密度和时间一致性的异常分数计算方法的计算成本不高，不会额外增加检测阶段的时间成本。

4.4.3 定性分析

在这一小节，我们对生成的结果进行定性分析，使用 Avenue 数据集对网络进行训练。我们在训练的过程中使用不同的损失函数，并对比使用不同损失函数时获得的预测帧画面，观察这两种损失函数能否提升预测帧的质量。

使用内容损失时得到的结果如 4-6所示。左图为原视频帧，右边的三个细节图分别对应着原视频帧，不使用内容损失的预测帧和使用内容损失的预测帧。细节图展示的监控画面右上角的两个人。我们可以看出：在女生肩膀的细节纹理部分，使用内容损失的预测帧能够对白色的高光进行刻画，比不使用内容损失的预测帧更接近原帧；在女生右下角与墙壁交界的地方，使用内容损失的预测帧获得的边界更直，更接近原帧，而不使用内容损失的预测帧在边界处较为模糊。由此可以看出，在原有损失的基础上增加内容损失可以提高预测帧的质量，使其更接近真实帧。

使用误差密度损失时得到的结果如图 4-7所示。左图为原视频帧，右边的三个细节图分别对应原帧，不使用误差密度损失的预测帧和使用误差密度损失的预测帧。细节图展示的是画面右侧行走的小朋友。比较三者的细节可以发

图 4-6: 内容损失的定性分析

现：在小朋友的衣服处，使用误差密度损失的预测帧对衣服褶皱和帽子的细节刻画的更为清晰，衣服的褶皱更立体，帽子显得更蓬松。使用误差密度损失可以使得预测帧在局部细节上更接近原帧。由定性结果分析可得，在训练阶段引入内容损失和误差密度损失，可以使训练的结果在内容和细节部分更接近原来的真实帧。

图 4-7: 误差密度损失的定性分析

4.4.4 消融实验

我们设计了一个消融实验，来证明两种损失和异常分数计算方法的有效性。当不使用某种损失时，只需要在训练阶段将对应的 $\lambda_{content}$ 或 $\lambda_{diffDensity}$ 设为 0 即可。消融实验的结果见表 4-2。表格的第一列对应的是内容损失函数，第二列对应的是误差密度损失函数，第三列对应的是基于误差密度和时间一致性的异常分数计算方法，最后一列是最终得到的 AUC。表格中第一行不使用任何策略；第二行仅在训练阶段加入内容损失；第三行使用内容损失和我们提出的异

表 4-2: 算法在 Avenue 数据集上的消融实验

ContentLoss	DiffDensityLoss	ScoreCal	AUC
			85.0
✓			85.7
✓		✓	87.5
	✓		85.6
	✓	✓	88.1
✓	✓		86.1
✓	✓	✓	88.3

常分数计算方法；第四行仅使用误差密度损失；第五行使用误差密度损失和异常分数计算方法；第六行使用内容损失和误差密度损失；第七行同时使用内容损失、误差密度损失和异常分数计算方法。我们可以看到：

1. 第二行和第四行的结果均比第一行好，因此不论是单独使用内容损失或误差密度损失均能提高算法的性能；
2. 第三行的结果比第二行好，第五行的结果比第四行好，这说明在使用内容损失或者误差密度损失的基础上使用基于误差密度和时间一致性相结合的异常分数计算方法能够获得更好的结果；
3. 第六行的结果比第一行、第二行和第四行好，表明同时使用内容损失和误差密度损失比不使用这两种误差好，也比单独使用内容损失或误差密度损失效果好；
4. 第七行的结果比第六行好，证明内容损失和误差密度损失在与基于误差密度和时间一致性的异常分数计算方法结合时能够获得更好的效果。

4.5 本章小结

本章设计了基于内容损失和误差密度损失的损失函数。当前常见的损失函数主要集中在对视频外观信息和运动信息的约束上，而对于视频的具体内容没有进行分析。在此基础上，我们基于迁移学习提出了基于内容的损失函数，该方法利用在 ImageNet 上预训练好的 ResNet34 对视频帧的内容特征进行提取，在训练过程对重构或预测帧进行内容上的约束。当前常见的损失函数也是针对

整张图像的差异和进行计算的，因此对小面积的细节关注不足，我们再次使用了差异密度的思想，把差异密度当做训练过程中的损失函数，目的是获得与当前实际帧在细节上更一致的输出。综合使用这两种损失函数进行网络的训练，可以获得质量更高的预测或者重构帧，进而提高视频异常检测算法的表现。最后我们对这两种损失函数进行了实验和消融实验，实验证明此方法与其他算法相比，能在 Avenue 数据集上取得具有竞争力的 AUC，定性分析说明该方法够获得质量更高的预测帧，显著提高视频异常检测的精度。

第五章　算法在煤矿智能视频分析系统中的应用

本章介绍了煤矿智能视频分析系统，并将提出的算法应用到系统中去，进行皮带异物的检测。该系统能够对皮带运输机上的煤流进行检测，看是否有异物出现，并及时向工作人员报警，具有很大的实际应用价值。

5.1　煤矿智能视频分析系统介绍

5.1.1　煤矿智能视频分析系统背景

随着监控摄像头的技术不断进步，相关的系统和平台也在不断发展，这些监控系统已经被广泛应用于各个场所，在加强安全，规范行为，方便追溯，提高效率等方面起到了至关重要的作用。

我国是能源生产大国也是能源消费大国，其中煤炭占能源消费总量的三分之二以上。因此，煤炭事业的发展成为经济建设的基础之一。2020 年 3 月，发展改革委、能源局、应急部、煤矿安监局、工业和信息化部、财政部、科技部和教育部联合印发了《关于加快煤矿智能化发展的指导意见》。意见提出，到 2021 年基本实现井下和露天煤矿固定岗位的远程监控，到 2035 年，各类煤矿基本实现智能化。

煤矿环境复杂，人员、设备、环境存在极大不确定性，传统基于人工监测的手段存在人力成本高、时效性差、无法量化管理的短板；通过智能视频分析系统平台可以实现全天候 24 小时不间断监测，对各项监测分析结果进行量化展示，加以人工决策判断，实现了危害感知、违规预警、事后溯源、人性化管理，有助于提高矿井安全生产，大大提高矿井智能化管理水平。

表 5-1: 煤矿智能视频分析系统的功能需求

功能	描述
首页监测功能	值班人员能够直观全面地看到各类智能视频分析情况及异常问题汇总并在告警第一时间做出反应
实时视频浏览	调度人员可根据筛选条件，快速准确找到指定地点视频并查看视频分析结果
分析规则设置	对指定地点指定摄像仪进行分析、告警规则设置
钻场作业录像	对钻场作业录像进行标识牌文字智能识别
皮带堆煤告警	够识别皮带转载点或其它指定地点堆煤情况，发生堆煤现象时系统发出告警
皮带异物告警	监测运输皮带上物体，对异物（非煤流）进行大小、形状进行识别，当识别到异物时系统平台立即发出告警
烟火告警	实时监测指定地点是否有明烟、明火，当识别监测出时，系统平台立即发出告警
禁区闯入告警	自定义危险区域及非危险区域，实时监测指定地点进行分析，当人员进入划定危险区域时，系统平台立即发出告警
工作面禁区闯入告警	采煤机运行期间人员闯入告警；采煤机端头割煤时三角区人员闯入告警；人员跨越刮板运输机告警；液压支架移架时人员架前作业告警；电缆槽人员站立告警；工作面刮板输送机缺刮板、断链告警；工作面刮板输送机机头堆煤告警；工作面液压支架护帮状态智能识别报警
摄像仪管理	添加删除摄像仪及进行摄像仪相关管理
预警中心	相关管理人员对触发告警规则的视频、图片、信息进行查阅及人工复核
用户管理	对使用人员账号及权限分配管理

5.1.2 煤矿智能视频分析系统需求

煤矿智能视频分析系统平台分为采集端、服务分析端、电脑终端和移动终端。平台支持多种报警方式（例如语音提醒、短信提醒、外接告警设备等），支持报警的搜索、联动、推送；支持与煤矿现有系统联动及数据推送接收，同时支持与第三方平台、告警设备等多种信号接入和输出。

该智能分析系统的功能需求见表5-1，主要是对井下的各种监控视频进行查看、管理、分析和报警。图5-1展示了表格中的工作面禁区闯入告警的部分

需求，是矿井下有人员闯入禁区的几种典型场景。左图是斜巷下料时，有人与车同行；中图是工作面采煤时，有人员闯入；右图是电缆槽里有人。这几种场景发生时会对生产的安全造成极大的隐患，极易引起生产事故。因此需要通过监控视频对这几个场景进行监控，判断是否有人员闯入禁区，并及时报警，进行相应的处理。

图 5-1: 煤矿监控视频中有人闯入禁区

在这些视频分析需求中，皮带异物告警场景下可以采用视频异常检测算法，通过对正常的皮带运输情况进行学习，建立正常情况下的模型，然后对监控视频进行检测，看皮带上是否出现非煤流的物体，或者比较大的煤块，并在发现异物和大煤块时进行报警。值班人员可根据报警信息进行后续的处理，判断是否需要停机，以避免异物对皮带运输系统造成损害，对生产安全造成不良影响。

5.2 算法在皮带异物告警场景下的应用

5.2.1 皮带异物告警需求

在煤炭的采掘、生产、转运和加工过程中，皮带输送机得到了广泛的使用。煤矿用皮带输送机可以在煤矿的复杂环境下进行大规模的煤炭运输，具有承载量大和运输距离长的特点。当皮带上有异物或者较大煤块时（如图5-2所示)，如不能进行相应的处理，轻则对皮带造成损伤，影响系统的正常运转，带来经济损失，重则带来安全隐患，导致严重的生产事故。因此，需要对皮带上运送的内容进行检测，看是否有异物或者大煤块。基于人工的方式需要较多人力资源，且具有不确定性。因此可以使用计算机进行智能检测。

在这一场景下，我们选择使用视频异常检测技术进行检测。不采用传统目

标识别方法的原因主要有两个，一是在实际生产过程中，我们只需要对异物和大煤块进行判断，并不需要检测出异物的具体类别。二是异物是多种多样的，且具有未知性，我们不知道皮带上会有何种异物出现。目标识别技术只能对有限种已知的物体进行识别，并不能满足上述要求。而使用视频异常检测技术可以有效的避免这些问题，高效地进行检测。

图 5-2: 皮带异物

在皮带异物告警场景下，主要的需求有两类，一类是功能需求，一类是性能需求。在功能方面，系统需要令监控摄像头对准皮带运输区域，并采集相应的视频数据。然后利用算法对视频数据进行分析，当画面中出现异物，如铁杆、杂物等时，或煤流中含有大煤块时，获得较高的异常分数，若该分数超过相关工作人员手动设置的阈值，则进行报警，并反馈给工作人员。在性能方面，系统要求算法能对视频进行实时处理，因此在检测阶段，模型对视频帧的处理速度要达到每秒 25 帧 (Frames Per Second，pfs) 以上。

5.2.2 算法设计

该算法主要包含两个部分，一个是训练部分，一个是检测部分。训练部分需要对正常情况下的数据进行采集，然后对采集到的训练数据进行预处理，再利用预处理得到的数据训练模型，获得能够对正常场景进行描述的模型。在检测部分，利用模型对获得的视频流进行处理，获得最终的异常分数。

算法的整体架构如图5-3所示。训练阶段分为三个步骤，正常视频采集，视频预处理和模型训练。首先是正常视频的采集，此阶段需要在煤矿下安装摄像头，拍摄固定角度的皮带运输场景。拍摄的视频要确保没有异常情况的发生，才能进行用于训练。若训练视频中包含异物或者大煤块，则学习时会将此类情况视为正常，学到的模型不能很好的分辨异常情况。因此，需要人为对视频进行筛选，确保训练视频中没有异常情况的发生。

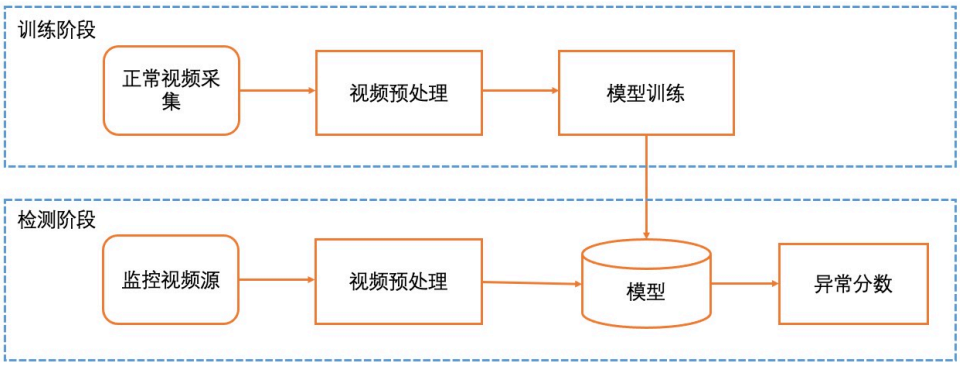


图 5-3: 算法框架

获取正常视频后，我们需要对视频进行预处理，将视频帧转化为 256×256 大小，并将其归一化到到 [0,1] 范围内。训练时使用的模型同4.4.1章，用 U-Net 作为基础的网络对视频帧进行预测，同时使用连续的 5 帧进行训练，以前 4 帧为输入，第 5 帧为输出，使得网络能够利用前 4 帧预测第 5 帧。U-Net 包括解码器和生成器两个部分，可以利用解码器提取视频图像前的特征，然后将获得的特征输入到生成器中，以此预测第 5 帧的图像。

训练过程中使用了强度损失（式4-1)、梯度损失（式4-2) 和光流损失（式4-3)。强度损失对视频帧外观进行约束，梯度损失对视频帧的边缘内容进行约束，光流损失对运动信息进行约束。除此之外，还使用了第4章提出的内容损失 (式4-6) 和误差密度损失（式4-7）。内容损失可以保证生成帧与原帧在内容上相似，而误差密度损失使得二者在小面积范围内接近，可以获得更好的预测结果。在训练时，我们借鉴了生成对抗网络的训练思想，以 U-Net 预测模块为生成器，利用 PatchGan 中的判别器判别是否为真实视频帧，二者相互对抗，训练至参数收敛。

检测阶段首先接入视频源，获取实时视频帧。然后对视频帧进行预处理，将视频帧转化为 256×256 大小，并进行归一化。接着将获得的视频帧 4 帧一组送到模型中，获得预测的帧，然后将预测的帧与实际的帧进行对比，使用第3章提出的基于误差密度和基于时间一致性的异常分数计算方法获得最终的异常值。由于基于误差密度和基于时间一致性的异常分数计算方法都不是在线计算方法，因此，我们对这两种方法进行了下面的改进：

在使用基于误差密度的异常分数计算方法时，需要用式3-7中的公式对最终的异常分数进行归一化，因此在计算过程中，需要对 f_t 的最大值和最小值

进行记录，当发现大于最大值或者小于最小值的 f_t 时，需要对记录进行实时更新。

在使用基于时间一致性的异常分数计算方法时，会参考前后 k 帧的差异情况对当前帧的异常程度进行判断。而在实时视频流中获取不到后 k 帧的结果，因此，我们修改了分数的计算方式，在计算第 i 帧的异常分数时，只参考前 k 帧的结果，不对后 k 帧进行计算。用式5-1取代式3-11对最终的异常分数进行计算。

$$final\_ano\_score_i = \frac{\sum_{n=i-k}^{i} w_{n,m} \times ano\_score_n}{\sum_{n=i-k}^{i} w_{n,m}} \tag{5-1}$$

5.2.3 算法实现

由于该系统尚未落地，不能和井下采集的实时视频进行连接，我们使用预先收集的井下视频进行效果验证。预先收集的视频包含两个视角的不同视频，一个是在皮带正上方俯视的视角，一个是在皮带前上方俯视的视角。皮带正上方俯视的视角共有 20 个视频片段，长度 3-5 分钟不等，累计 73 分钟，其中有 3 个视频片段含有异物。我们将其分为 15 个训练视频，5 个测试视频，训练视频均为正常视频，测试视频中有 3 个片段为异常视频，两个片段为正常视频。皮带前上方俯视的视角共有 14 个视频片段，长度 3-6 分钟不等，累计 68 分钟，其中有 3 个视频含有异物。我们将其分为 11 个训练视频，3 个测试视频，训练视频均为正常视频，测试视频中有 3 个片段为异常视频，1 个片段为正常视频。通过对这两个视角的视频进行学习，获得两个不同的异常检测模型，然后用模型对同一视角的视频进行异常检测。

在训练的过程中，生成器的学习率为 0.0002，判别器的学习率为 0.00002，$\beta_1 = 0.9$，$\beta_2 = 0.999$，batch 大小为 4，训练 1000 个 epoch。输入序列的长度 $p = 5$。损失函数的参数为：$\lambda_{int} = 2$，$\lambda_{gd} = 2$，$\lambda_{op} = 2$，$\lambda_{adv} = 0.05$，$\lambda_{content} = 1$，$\lambda_{diffDensity} = 2$。在检测过程中，我们采用了基于误差密度和时间一致性的异常分数计算方法。其中平均池化的核大小为 12，时间一致性的 $k = 11$，$\theta = 9$。

算法使用 Python 语言实现。在图像的预处理和最终结果的展示阶段使用了 cv2 和 numpy 库，网络的训练阶段使用了 pytorch 库，内容损失中预训练的 ResNet 模型使用了 torchvision.models 模块的 resnet34 类，除此之外还用了 sys，scipy，os 等库进行辅助的操作。算法在 NVIDIA 1080-Ti GPU 上运行。

5.2.4 算法效果

首先在不含异物的场景下进行检测，检测的结果如图5-4所示。摄像机在皮带正上方对皮带进行俯拍，获得皮带上运输煤流的情况。此段视频中皮带上的煤流在稳定运输，没有异常发生。将该段视频输入到算法的检测阶段，得到的结果如图5-5所示，整体的异常分数都比较低。异常分数越高，发生异常的可能性越大。而这段视频的异常分数集均在较低的范围内，说明整段视频大概率没有异常发生。

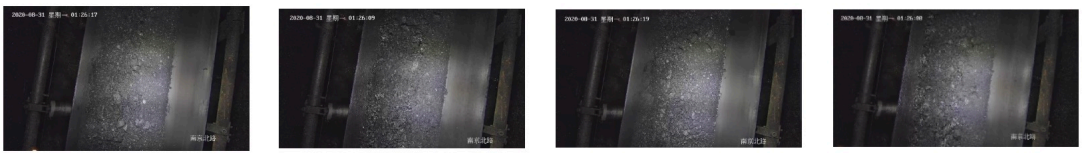


图 5-4: 皮带上没有异物或大煤块

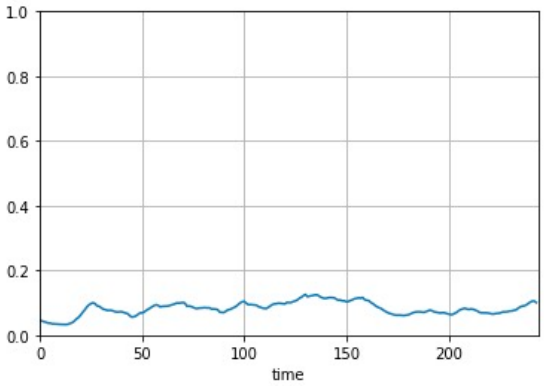


图 5-5: 皮带上没有异物或大煤块时的检测结果

接着我们对含有异常的场景进行检测，检测的场景如图5-6所示，此时摄像机在皮带前上方对皮带进行俯拍。在该视频中一段时间内，皮带上出现了大煤块，如图中红色框中所示。将该段视频输入到我们提出的算法中进行检测，得到的结果如图5-7所示，可以看出，当大煤块出现时，异常分数明显升高，表明此时发生异常的概率比较大。

在运行算法时，检测速度是25fps，能够达到实时处理的标准。利用算法对煤块皮带视频进行检测，可以获得对应的异常分数。相关工作人员可以在视频分析系统内进行告警规则设置，设定告警阈值，当异常分数超过阈值时则产生警报。

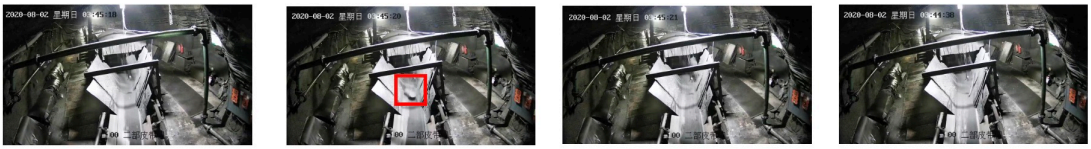


图 5-6: 皮带上有大煤块

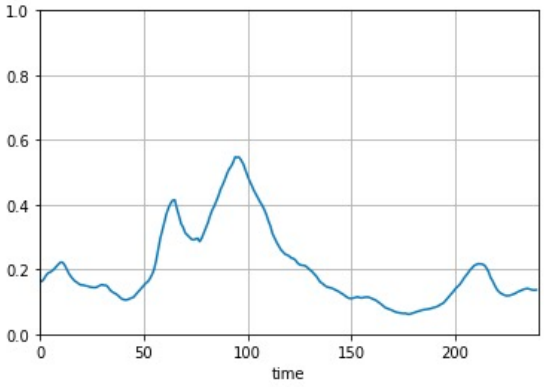


图 5-7: 皮带上有大煤块时的检测结果

5.3 本章小结

本章介绍了将提出的视频异常检测算法应用于煤矿智能视频分析系统中的案例。使用本文提出算法可以有效的对皮带上的异物和大煤块进行检测，避免异物或大煤块对皮带造成影响，给煤矿的生产安全带来危害。通过在煤矿智能分析系统中的应用，充分验证了本文提出算法的有效性和实用性，证明该算法能够应用于实际生产生活。

第六章　总结与展望

6.1　总结

随着监控摄像头在各行各业的广泛应用，对摄像头拍摄内容的分析与处理也成了人们关注的问题。视频异常检测可以利用算法自动地检测监控视频中的异常，发出警报，方便相关工作人员进行后续的处理。然而，视频异常检测任务中仍存在着一些问题，如：小面积异常的漏检问题，正常情况的不可预测性问题和监控视频画面突变的问题。为了克服这些问题，我们提出了基于误差密度和时间一致性的异常分数计算方法。除此之外，现有视频异常检测算法在重构或者预测时很少关注视频画面的具体内容，也没有对局部细节进行很好的约束。为此我们提出了一种基于内容损失和误差密度损失的损失函数。本文的主要贡献如下:

1. 本文提出了一种基于误差密度和时间一致性的异常分数计算方法。基于误差密度的异常检测算法参考了卷积神经网络中平均池化的思想，对差异图进行平均池化后求最大值，可以对小面积的异常进行响应。基于时间一致性的异常分数计算方法用多帧的误差代替一帧的误差进行计算，能够有效缓解正常情况的不可预测性问题和监控视频画面突变的问题，也能降低噪声对检测结果的影响。通过实验，本文论证了两个方法的有效性，并证明现有算法与两个方法结合后能够显著提升检测水平。
2. 本文提出了一种基于内容损失和误差密度的损失函数。基于内容损失的损失函数使用了迁移学习的思想，利用预训练好的模型对视频帧进行特征提取，获得视频内容的表示，并在原视频帧和预测或重构帧的内容上进行约束。基于误差密度的损失函数使用平均池化的思想对局部的误差加以约束，保证两帧之间的局部差异不会过大。通过实验和消融实验证明，这两种损失函数能够对视频的内容和局部差异进行约束，并能与基于误差密度和时间一致性的异常分数计算方法相结合，显著提高算法的异常检测水平。
3. 本文将提出的方法应用到煤矿智能视频分析系统中，能够利用异常检测算

法对皮带负载情况进行检测，在有异物或者大煤块的时候发生报警，为保证生产安全和提高生产效率起到了重要作用。

6.2 展望

近几年，监控视频数据变得越来越多，且对异常处理的实时性要求越来越高。但是当前的研究主要集中在检测水平的改进上，算法的实时性问题和增量性问题还没有被充分的研究。因此，本文建议通过这两个方面对视频异常检测技术进行深入研究：一是视频异常检测的在线学习。不仅在检测阶段，在训练阶段也能够进行在线学习。通过在线学习的方法，能够实时的对视频流进行处理，获得异常信息，及时发出报警；二是视频异常检测的增量学习。通过增量学习的方法，对摄像头内新出现的情况进行增量的学习，在不遗忘旧知识的情况下对新的场景进行学习，使得当摄像头拍摄的场景发生变化时，算法也能很快的学到新的内容，可以更好地应用在云台等可旋转摄像头上。

参考文献

[1] 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.

[2] CHALAPATHY R, CHAWLA S. Deep learning for anomaly detection: A survey[J]. arXiv preprint arXiv:1901.03407, 2019.

[3] LIU F T, TING K M, ZHOU Z-H. Isolation forest[C] // 2008 eighth ieee international conference on data mining. 2008 : 413 – 422.

[4] NGUYEN T-N, MEUNIER J. Anomaly detection in video sequence with appearance-motion correspondence[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019 : 1273 – 1283.

[5] 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.

[6] MORAIS R, LE V, TRAN T, et al. Learning regularity in skeleton trajectories for anomaly detection in videos[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019 : 11996 – 12004.

[7] PARK H, NOH J, HAM B. Learning memory-guided normality for anomaly detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 : 14372 – 14381.

[8] CHANG Y, TU Z, XIE W, et al. Clustering Driven Deep Autoencoder for Video Anomaly Detection[C] // European Conference on Computer Vision. 2020 : 329 – 345.

[9] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 : 2921 – 2929.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 : 770 – 778.

[11] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018 : 6479 – 6488.

[12] RAMACHANDRA B, JONES M. Street Scene: A new dataset and evaluation protocol for video anomaly detection[C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020 : 2569 – 2578.

[13] KIRAN B R, THOMAS D M, PARAKKAL R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos[J]. Journal of Imaging, 2018, 4(2) : 36.

[14] 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.

[15] LI W, MAHADEVAN V, VASCONCELOS N. Anomaly detection and localization in crowded scenes[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(1) : 18 – 32.

[16] ADAM A, RIVLIN E, SHIMSHONI I, et al. Robust real-time unusual event detection using multiple fixed-location monitors[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(3) : 555 – 560.

[17] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C] // 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) : Vol 1. 2005 : 886 – 893.

[18] XU D, SONG R, WU X, et al. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts[J]. Neurocomputing, 2014, 143 : 144 – 152.

[19] GIBSON J J. The perception of the visual world.[J], 1950.

[20] CONG Y, YUAN J, LIU J. Sparse reconstruction cost for abnormal event detection[C] // CVPR 2011. 2011 : 3449 – 3456.

[21] WU S, MOORE B E, SHAH M. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes[C] // 2010 IEEE computer society conference on computer vision and pattern recognition. 2010 : 2054 – 2060.

[22] 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.

[23] 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.

[24] CHENG K-W, CHEN Y-T, FANG W-H. Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation[J]. IEEE Transactions on Image Processing, 2015, 24(12) : 5288 – 5301.

[25] ZHANG Y, LU H, ZHANG L, et al. Combining motion and appearance cues for anomaly detection[J]. Pattern Recognition, 2016, 51 : 443 – 452.

[26] YANG X, LATECKI L J, POKRAJAC D. Outlier detection with globally optimal exemplar-based GMM[C] // Proceedings of the 2009 SIAM International Conference on Data Mining. 2009 : 145 – 154.

[27] SATMAN M H. A new algorithm for detecting outliers in linear regression[J]. International Journal of statistics and Probability, 2013, 2(3) : 101.

[28] LATECKI L J, LAZAREVIC A, POKRAJAC D. Outlier detection with kernel density functions[C] // International Workshop on Machine Learning and Data Mining in Pattern Recognition. 2007 : 61 – 75.

[29] GAO J, HU W, ZHANG Z M, et al. RKOF: robust kernel-based local outlier detection[C] // Pacific-Asia conference on knowledge discovery and data mining. 2011 : 270 – 283.

[30] HIDO S, TSUBOI Y, KASHIMA H, et al. Statistical outlier detection using direct density ratio estimation[J]. Knowledge and information systems, 2011, 26(2) : 309 – 336.

[31] BREUNIG M M, KRIEGEL H-P, NG R T, et al. LOF: identifying density-based local outliers[C] // Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000 : 93 – 104.

[32] TANG J, CHEN Z, FU A W-C, et al. Enhancing effectiveness of outlier detections for low density patterns[C] // Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2002 : 535 – 548.

[33] JIN W, TUNG A K, HAN J, et al. Ranking outliers using symmetric neighborhood relationship[C] // Pacific-Asia conference on knowledge discovery and data mining. 2006 : 577 – 593.

[34] KRIEGEL H-P, KRÖGER P, SCHUBERT E, et al. LoOP: local outlier probabilities[C] // Proceedings of the 18th ACM conference on Information and knowledge management. 2009 : 1649 – 1652.

[35] TANG B, HE H. A local density-based approach for outlier detection[J]. Neurocomputing, 2017, 241 : 171 – 180.

[36] DANG T T, NGAN H, WEI L. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data[C] // IEEE International Conference on Digital Signal Processing. 2015.

[37] KNORR E M, NG R, TUCAKOV V. Distance-based outliers: algorithms and applications[J]. The VLDB Journal —The International Journal on Very Large Data Bases, 2000.

[38] RAMASWAMY S, RASTOGI R, SHIM K, et al. Efficient Algorithms for Mining Outliers from Large Data Sets[J]. ACM, 2000.

[39] ANGIULLI F, BASTA S, PIZZUTI C. Distance-based detection and prediction of outliers[J]. IEEE transactions on knowledge and data engineering, 2005, 18(2): 145 – 160.

[40] HE Z, XU X, DENG S. Discovering cluster-based local outliers[J]. Pattern Recognition Letters, 2003, 24(9-10): 1641 – 1650.

[41] DING Z, FEI M. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window[J]. IFAC Proceedings Volumes, 2013, 46(20): 12 – 17.

[42] XU D, WANG Y, MENG Y, et al. An improved data anomaly detection method based on isolation forest[C] // 2017 10th International Symposium on Computational Intelligence and Design (ISCID): Vol 2. 2017: 287 – 291.

[43] CAMPOS G O, ZIMEK A, MEIRA W. An unsupervised boosting strategy for outlier detection ensembles[C] // Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018: 564 – 576.

[44] ABE N, ZADROZNY B, LANGFORD J. Outlier detection by active learning[C] // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 504 – 509.

[45] XU D, RICCI E, YAN Y, et al. Learning deep representations of appearance and motion for anomalous event detection[J]. arXiv preprint arXiv:1510.01553, 2015.

[46] LUO W, LIU W, GAO S. Remembering history with convolutional lstm for anomaly detection[C] // 2017 IEEE International Conference on Multimedia and Expo (ICME). 2017: 439 – 444.

[47] 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. 2015: 234 – 241.

[48] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. arXiv preprint arXiv:1406.2661, 2014.

[49] 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. 2020 : 14183 – 14193.

[50] MEDEL J R, SAVAKIS A. Anomaly detection in video using predictive convolutional long short-term memory networks[J]. arXiv preprint arXiv:1612.00390, 2016.

[51] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[52] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutional networks[C] // Proceedings of the IEEE international conference on computer vision. 2015 : 2758 – 2766.

[53] RODRIGUES R, BHARGAVA N, VELMURUGAN R, et al. Multi-timescale trajectory prediction for abnormal human activity detection[C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020 : 2626 – 2634.

[54] 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. 2017 : 1933 – 1941.

[55] IONESCU R T, SMEUREANU S, POPESCU M, et al. Detecting abnormal events in video using narrowed normality clusters[C] // 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). 2019 : 1951 – 1960.

[56] VU H, NGUYEN T D, LE T, et al. Robust anomaly detection in videos using multilevel representations[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 33. 2019 : 5216 – 5223.

[57] YU G, WANG S, CAI Z, et al. Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events[C] // Proceedings of the 28th ACM International Conference on Multimedia. 2020 : 583 – 591.

[58] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[59] LAI W-S, HUANG J-B, WANG O, et al. Learning blind video temporal consistency[C] // Proceedings of the European conference on computer vision (ECCV). 2018: 170–185.

[60] GONG W, WANG W, LI W, et al. Temporal consistency based method for blind video deblurring[C] // 2014 22nd International Conference on Pattern Recognition. 2014: 861–864.

[61] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C] // Proceedings of the IEEE international conference on computer vision. 2015: 4489–4497.

[62] 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. 2017: 341–349.

[63] GONG D, LIU L, 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. 2019: 1705–1714.

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

[65] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks[C/OL] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. http://lmb.informatik.uni-freiburg.de//Publications/2017/IMKDB17.

[66] FAN L, HUANG W, GAN C, et al. End-to-end learning of motion representation for video understanding[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6016–6025.

[67] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 3431 – 3440.

[68] DU X, LIN T-Y, JIN P, et al. Spinenet: Learning scale-permuted backbone for recognition and localization[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 : 11592 – 11601.

[69] BEERY S, WU G, RATHOD V, et al. Context r-cnn: Long term temporal context for per-camera object detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 : 13075 – 13085.

[70] ISOLA P, ZHU J-Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 : 1125 – 1134.

[71] TAN C, SUN F, KONG T, et al. A survey on deep transfer learning[C] // International conference on artificial neural networks. 2018 : 270 – 279.

[72] 邱锡鹏. 神经网络与深度学习 [M/OL]. 北京 : 机械工业出版社, 2020. https://nndl.github.io/.

[73] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[J]. arXiv preprint arXiv:1411.1792, 2014.

[74] DENG J, DONG W, SOCHER R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C] // CVPR09. 2009.

[75] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25 : 1097 – 1105.

[76] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[77] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 1 – 9.

[78] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

[79] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[80] TAN M, CHEN B, PANG R, et al. Mnasnet: Platform-aware neural architecture search for mobile[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019 : 2820 – 2828.

[81] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C] // International Conference on Machine Learning. 2019 : 6105 – 6114.

[82] KOLESNIKOV A, BEYER L, ZHAI X, et al. Big transfer (bit): General visual representation learning[J]. arXiv preprint arXiv:1912.11370, 2019, 6(2) : 8.

[83] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: algorithms and applications[J]. The VLDB Journal, 2000, 8(3) : 237 – 253.

[84] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using lstms[C] // International conference on machine learning. 2015 : 843 – 852.

[85] LIU S W, NGAN H Y, NG M K, et al. Accumulated relative density outlier detection for large scale traffic data[J]. Electronic Imaging, 2018, 2018(9) : 239 – 1.

致　　谢

转眼间三年的研究生生活即将结束，很庆幸来到 RINC 这个大家庭，遇到了非常好的导师和非常优秀的同学，在这里度过了充实而又美好的时光。

首先，我要感谢我的导师申富饶教授。在三年的研究生生活中，申老师给予了我太多的帮助。刚入学时，我从对这个方向了解的比较少，是申老师一步步地引导我逐渐走上正轨。申老师每周都会与我们讨论，在讨论的过程中让我对科研中的问题有了更为深入的理解。其次，我还要感谢赵健老师。赵健老师在组会时提出了很多具有启发性的问题，并在论文的写作方面颇有造诣，曾细致地帮我修改论文，令我受益良多。我还要感谢实验室的同学们，大家都是非常优秀且友好的人，我从大家的身上学到了很多很多。除此之外，我还要感谢我的家人们，是他们给了我无私的爱与帮助，也给我带来了快乐与希望。

简历与科研成果

基本信息

邵玥，女，汉族，1993 年 6 月出生，安徽宿州人。

教育背景

2018 年 9 月 — 2021 年 6 月 南京大学计算机科学与技术系 硕士
2012 年 9 月 — 2016 年 6 月 南京航空航天大学计算机科学与技术学院 本科

攻读硕士学位期间的专利成果

1. 申富饶，邵玥，姜少魁，刘凤山，金勇，盛敏。一种基于激光雷达的煤矿刮板机负载高度检测的方法 (202010321757.6)
2. 申富饶，刘凤山，邵玥，金勇，盛敏。一种基于激光雷达的液压支架对齐方法 (202010326088.1)

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

2021 年 5 月 20 日

论文题名	基于深度学习的视频异常检测研究				
研究生学号	MP1833023	所在院系	计算机科学与技术系	学位年度	2021
论文级别	☑ 硕士　☐ 硕士专业学位 ☐ 博士　☐ 博士专业学位　（请在方框内画勾）				
作者 Email	shaoyue1993@163.com				
导师姓名	申富饶				

论文涉密情况：

☑ 不保密

☐ 保密，保密期：_____年_____月_____日 至 _____年_____月_____日

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。