

# 上下文建模与推理的视频异常事件检测

孙澈<sup>1)</sup> 武玉伟<sup>2),1)</sup> 贾云得<sup>2),1)</sup>

<sup>1)</sup>(北京理工大学计算机学院智能信息技术北京市重点实验室 北京 100081)

<sup>2)</sup>(深圳北理莫斯科大学广东省智能感知与计算重点实验室 广东 深圳 518172)

**摘要** 视频异常事件检测旨在从视频中自动地检测出不符合正常事件规律的视频事件。视频中许多正常和异常的事件是由目标与场景或其它目标交互而产生的,即它们是以目标为中心且高度上下文相关的。如何从底层的视频特征中提取事件高层语义上下文信息,并根据上下文信息进行视频异常事件检测仍是一个开放的难题。为此,本文提出了一种新的上下文建模与推理的视频异常事件检测方法。本文方法通过建立视频的上下文图,自动地推理事件相关的语义上下文信息,以缩小底层视觉特征与异常事件高层语义之间的差距,实现异常事件检测。具体来说,首先使用了预训练的目标检测网络,提取目标初始的表观特征、目标之间的时空关系特征和场景特征;其次设计了一个上下文图推理模块,通过建模时空上下文图,将提取到的特征显式地建模为三类语义上下文,包括事件目标的个体行为、不同目标之间的时空关系以及目标与场景之间的交互,其中图的节点表示目标/场景,图的边表示时空关系;最后构建了一个异常预测模块,根据推理到的语义上下文信息进行异常事件检测。本文的上下文图推理模块基于平均场理论,通过使用多个带有消息传递模块的循环神经网络,迭代更新图的节点和边的状态,目的是从底层的视觉特征中推理得到高层的语义上下文。本文的异常预测模块包括注意力池化网络层和全连接网络层,通过输入语义上下文信息,计算视频帧的异常分数,从而正确地进行异常事件检测。实验中,设计了一个自训练策略,分别使用了无监督、半监督、弱监督和监督四种训练策略,以端到端的方式训练时空上下文图推理模块和异常预测模块。本文方法在四个公开的数据集上进行了实验,包括三个半监督的数据集 Subway (Entrance/Exit)、Avenue 和 ShanghaiTech,以及一个监督的数据集 UCF-Crime。与不使用上下文的方法相比,本文方法在 Subway (Entrance/Exit)、Avenue 和 ShanghaiTech 数据集上的无监督 AUC 指标分别提高了 2.7%/3.1%、2.0% 和 2.9%,半监督 AUC 指标分别提高了 3.5%/3.3%、4.0% 和 4.3%。在监督数据集 UCF-Crime 上,与没有使用上下文的方法相比,本文方法在半监督 AUC、弱监督 AUC 和监督 AUC 的指标上分别提高了 2.1%、0.4% 和 9.2%,取得了有竞争力的表现。

**关键词** 异常事件检测;上下文建模与推理;上下文图;自训练策略;深度学习

**中图法分类号** TP18 **DOI号** 10.11897/SP.J.1016.2024.02368

## Context Modeling and Reasoning for Video Abnormal Event Detection

SUN Che<sup>1)</sup> WU Yu-Wei<sup>2),1)</sup> JIA Yun-De<sup>2),1)</sup>

<sup>1)</sup>(Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology,

Beijing Institute of Technology, Beijing 100081)

<sup>2)</sup>(Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen, Guangdong 518172)

**Abstract** Video abnormal event detection aims to automatically detect events that do not conform to the regularities of normal events in videos. Many normal events and abnormal events in videos are caused by the interactions between event objects and scenes or other objects, and thus they are usually object-centric and highly contextual. Currently, it is still an open problem

收稿日期:2023-10-21;在线发布日期:2024-07-03。本课题得到深圳市自然科学基金面上项目(JCYJ20230807142703006)和广东省教育厅普通高校重点科研平台和项目(2023ZDZX1034)资助。孙澈,博士,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、机器学习和视频分析。E-mail: sunche@bit.edu.cn。武玉伟(通信作者),博士,长聘副教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、机器学习、多媒体分析。E-mail: wuyuwe@bit.edu.cn。贾云得,博士,教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、人工智能、认知计算和系统。

to discriminate abnormal events by acquiring high-level semantic context information from low-level visual features in videos. To this end, we propose a novel context modeling and reasoning method for video abnormal event detection. The method mines event-related semantic context information from video data by generating video context graphs, which is able to narrow the semantic gap between the low-level visual features in videos and the high-level semantics of abnormal events, and then uses the semantic context information to discriminate abnormal events correctly in videos. Specifically, we first use a pre-trained object detection neural network to extract the initial appearance features of all objects, the spatio-temporal relationship features between different objects, as well as the scene features. Then we devise a context graph inference module to explicitly model three types of semantic contexts, including individual object behaviors, pairwise relationships among different objects, and interactions between objects and scenes, where the nodes of the graph could describe the object and scene features, and the edges of the graph describe the spatio-temporal relationship features. We finally build an anomaly prediction module to discriminate abnormal events according to the semantic contexts captured from the previous context graph in videos. The proposed context graph inference module is based on the mean-field theory, and includes multiple recurrent neural networks with message-passing modules. The message-passing modules iteratively update the state of nodes and edges in the context graph for inferring the high-level semantic contexts from the low-level feature representations. The proposed anomaly prediction module consists of two attention-pooling network layers and one fully-connected network layer. The obtained context information is finally fed into the anomaly prediction module to calculate anomaly scores of all video frames for video abnormal event detection. In experiments, we introduce a self-training strategy to train the network models in four manners, including unsupervised, semi-supervised, weakly supervised and supervised manners. In this way, the spatio-temporal context graph inference module and anomaly prediction module are trained in an end-to-end manner seamlessly, such that they reinforce each other. The context reasoning method is evaluated on four public challenging datasets, including three semi-supervised datasets, i. e., the Subway (Entrance/Exit) dataset, Avenue dataset and ShanghaiTech dataset, as well as a supervised UCF-Crime dataset, respectively. Compared with existing methods without considering context modeling and reasoning, our context modeling and reasoning method improves the unsupervised *AUC* values by 2.7%/3.1%, 2.0% and 2.9% on the Subway (Entrance/Exit) dataset, Avenue dataset and ShanghaiTech dataset, and improves the semi-supervised *AUC* values by 3.5%/3.3%, 4.0% and 4.3%, respectively. Compared with existing methods without considering context modeling and reasoning on the supervised UCF-Crime dataset, our method significantly improves the semi-supervised, weakly-supervised and supervised *AUC* values by 2.1%, 0.4% and 9.2%, respectively.

**Keywords** abnormal event detection; context modeling and reasoning; context graph; self-training strategy; deep learning

## 1 引言

智能视频异常事件检测(video abnormal event detection)在交通监控、社会安防等诸多领域中受到了广泛关注。视频异常事件检测旨在运用先进的机

器学习和计算机视觉算法,实现自动地检测监控视频中偏离常规的事件或行为。由于监控场景的复杂性和人为定义规则的多样性,视频异常事件的界定通常呈现一定的模糊性(即“违反常规情况的事件”)<sup>[1]</sup>。这意味着待检测的异常事件不仅种类繁多且难以预设具体类型,例如在界定人群奔跑行为是

否异常时,需要考虑事件的背景和研究的范围,在街道上的奔跑行为通常被视为异常,而在马拉松赛事中同一奔跑行为则是正常的. 本文聚焦于社区、商场、学校人行道、道路交通、地铁进出站口等多种公共场所的安全监控视频中的异常事件检测任务. 在本文关注的事件背景和研究范围中,典型的异常事件涵盖了多种违法活动(如抢劫、打架、偷窃等)、交通违规事件(如人行道上驾驶摩托车和自行车、马路上危险嬉戏等)以及扰乱秩序的事件(如地铁站抛掷物品、逃票等). 总结起来,本文关注检测上述可能对社会公共安全构成潜在威胁的各种异常事件.

由于视频事件的多样性、异常事件的罕见性和高度上下文相关性,检测视频中的异常事件仍是一个热点研究课题<sup>[2]</sup>. 目前的大多数方法<sup>[3-6]</sup>利用了深

度神经网络来学习正常事件的时空模式,并将偏离正常模式的事件判定为异常事件. 这些方法在检测上下文无关的异常事件方面取得了显著成果. 然而,在现实世界中,许多异常事件是由目标与场景或其它目标交互而产生的,因此它们不仅以目标为中心,而且高度依赖于上下文关系. 现有的方法在检测这类异常事件时往往表现不佳,因为它们缺乏对视频事件视觉上下文的建模,即它们未充分考虑到视频中各个事件目标、它们之间的关系以及围绕事件的场景等视觉上下文信息. 以图 1 所示的车辆掉头事件为例,车辆在交通路口掉头是正常行为,但在高速公路上掉头则属于异常行为. 现有的方法往往忽略了交通路口和高速公路的上下文信息,这会导致错误的检测结果.

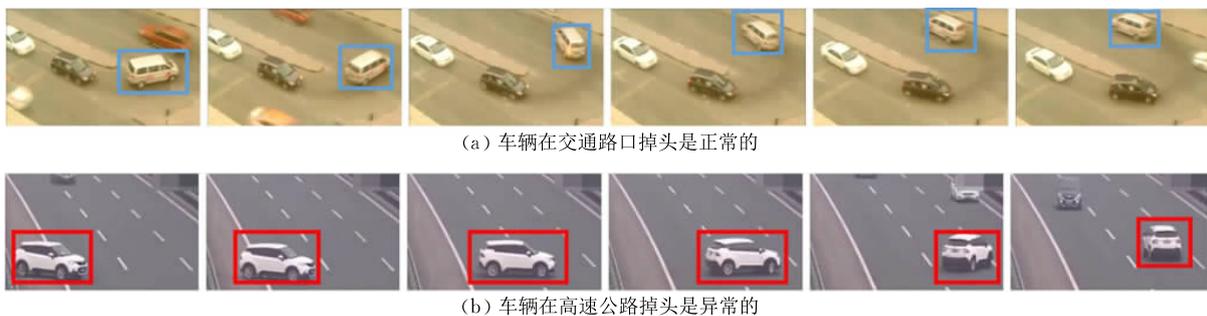


图 1 车辆掉头事件示例图

为此,最近的研究<sup>[7-9]</sup>引入了上下文信息来检测与上下文相关的异常事件,以提高异常检测的效果. 然而,这些方法在建模上下文信息时通常需要预先定义所有可能的语义上下文集合. 通常认为,事先定义所有上下文集合的正确性和完整性是难以保证的. 正如文献<sup>[10]</sup>所指出,在许多情况下,与上下文密切相关事件的定义具有多样性、不断变化性和不可预测性. 这意味着在这些方法中预先定义的上下文集合无法保证覆盖所有可能的异常事件,从而影响检测的准确度. 不同于这些方法,本文提出了一种上下文建模与推理的视频异常事件检测方法. 该方法通过生成和推理上下文图,从数据中自动建模和推理而不是手工定义上下文内容,以在不同场景中提高检测各类未知异常事件的准确度.

上下文建模与推理的视频异常事件检测方法的核心在于通过从视频中挖掘事件相关的语义上下文,缩小底层视觉特征与异常事件高层语义之间的语义差距. 本文方法通过上下文图推理得到了语义上下文,包括目标的个体行为、目标之间的成对关系以及目标和场景之间的交互,这些信息直接影响着视频事件的理解方式. 此外,本文方法还将上下文图

推理无缝地集成到异常预测任务中,使它们二者相互加强. 具体来说,本文方法构建了一个时空上下文图推理模块来捕获语义上下文,并设计了一个异常预测模块来预测异常分数. 本文方法首先使用预训练的区域备选网络和结构循环神经网络,从视频中提取目标、时空关系和场景的底层特征. 然后建立时空上下文图,其中图的节点表示目标/场景,图的边表示时空关系. 受到平均场(mean-field)理论的启发,本文通过使用多个带有消息传递模块的循环神经网络迭代更新图的节点和边的状态,进行语义上下文推理,可以从底层的视觉特征中推理得到高层的语义上下文. 异常预测模块通过使用两个独立的注意力池化网络层和一个全连接网络层,聚合推理得到语义上下文,最后计算帧级异常分数.

在实验中,本文方法设计了自训练(self-training)的策略,以端到端的方式同时训练特征提取模块、时空上下文图推理模块和异常预测模块. 由于缺少异常数据的标注,直接进行端到端的训练是很困难的. 为此,本文根据四种常见的异常标注情景设计了相应的训练方法,四种标注情景包括:(1)没有标注(无监督);(2)只有正常样本的标注(半监督);(3)有

粗粒度视频级标注(弱监督);(4)有细粒度帧级标注(监督).在这四种情景下,本文引入了一种统一的自训练策略完成端到端的网络优化.首先从无监督、半监督或弱监督的标注数据中获取初始化的帧级伪异常标注;然后使用伪标注数据训练网络;接下来根据预测结果优化更新伪异常标注.本文方法通过多次迭代优化伪标注和网络训练,可以获得理想的检测结果.本文在四个公开数据集UCF-Crime、Avenue、Subway和ShanghaiTech上进行了帧级异常评估.与最先进的方法相比,本文的方法在所有的情景下均取得了有竞争力的表现;与不使用上下文建模与推理的方法相比,本文的方法取得了很大的性能提升.

本文是已有文献[11]的扩展版本.文献[11]的贡献包括:(1)第一次提出了用于异常事件检测的上下文推理方法,能够从视觉特征中自动挖掘上下文信息,以缩小底层视觉上下文与异常事件高层语义之间的语义差距;(2)构建了一个用于表示和推理上下文的时空上下文图,充分利用了上下文信息来区分异常事件;(3)设计了一个基于图的深度高斯混合模型进行场景聚类,并采用半监督训练策略(此处特指训练集中仅包含正常数据的情况),以实现正常模式的有效学习.

本文对先前文献[11]进行了深入拓展,做出了以下创新和贡献:(1)相较于文献[11]采用场景聚类方法将视频场景划分为固定类别并在这些预设的场景类别上进行异常检测,本文创新性地场景信息整合至上下文图模型中.通过设计图推理方法从数据中自动学习并捕获场景自身的内在特征以及目标与场景间的交互特性,从而实现了从特定场景下的异常事件检测向更为普适的场景扩展;(2)在整合目标和时空关系表示方面,文献[11]直接对所有目标和时空关系进行异常评分,并将最高的异常分数作为视频帧的异常分数,而该策略易受视频上下文中的噪声干扰.相比之下,本文提出了一种基于注意力机制的异常预测模块,该模块能够有效地聚合与异常事件相关的目标、时空关系以及场景的上下文表示,从而更好地生成视频帧的异常分数;(3)在视觉特征提取方面,文献[11]依赖于预训练网络提取的底层视觉特征,在上下文推理过程中无法直接从原始视频像素中提取与事件紧密相关的语义上下文特征.而本文引入了一种新的自监督训练策略,使得特征提取模块、上下文图推理模块以及异常预测模块能够以端到端的方式联合优化,得以直接从视

频像素中提取最能体现异常事件特性的语义上下文特征.此外,这种自训练策略能够适应无监督、半监督乃至弱监督等不同标注条件下的异常检测任务,提升了模型的灵活性与有效性;(4)实验方面,本文不仅将文献[11]的方法推广到了多种标注数据受限的异常检测情境中,还进一步扩充了实验数据集,新增了Subway数据集上的实验结果.同时,本文增加了更多细致的消融实验分析,从而为本文所提出的改进策略提供了坚实的实验支持.

## 2 相关工作

### 2.1 视频异常事件检测

视频中的异常事件检测任务是计算机视觉和机器学习领域的研究热点.现有的研究方法<sup>[12-14]</sup>大致分为两类,一类是基于分类的方法,另一类是基于重建的方法.

(1)大多数基于分类的方法先提取视频特征,然后通过训练分类器对正常和异常数据进行分类.近年来,研究人员选择了使用深度网络提取视频特征,用深度特征取代传统手工提取的特征,他们将异常检测视为二分类问题(即正常和异常两个类别).由于深度学习需要大量的训练数据,而异常事件的罕见性使得获取足够标注的异常数据是很困难的,因此许多深度方法使用了仅由正常数据训练的单类分类器(one-class classifier)检测异常事件<sup>[15-17]</sup>.Ionescu等人<sup>[5]</sup>进一步创新性地设计了一个一对多分类器(one-versus-rest classifier),改进了单类分类器的分类效果,得到了更有区分性的分类边界.这些方法中的特征提取过程和异常分类预测过程是独立进行的.因此,学习到的特征难以保证与异常预测是相关的,可能导致次优的结果.与这些方法不同,本文将上下文图推理和基于分类的异常预测集成到了端到端的网络中,确保学习到事件相关的上下文特征表示.此外,不同于已有的单类分类器,本文设计了一个新的三元组(triplet)分类器,目的是减少正常数据的类内距离以及增加正常数据和异常数据之间的类间距离,从而提高分类器的判别能力.

(2)基于重建的方法学习正常事件模式,并通过重建误差识别异常事件.这类方法使用正常数据训练重建模型,期望模型能很好地重建正常数据而无法很好地重建异常数据.自编码器(auto-encoder)是一个常见的重建模型,已广泛地应用于异常事件检测.例如全卷积自编码器<sup>[4]</sup>、时空卷积自编码

器<sup>[6,18-19]</sup>、记忆自编码器<sup>[20]</sup>等.最近的方法<sup>[21]</sup>改进了重建模型,并通过预测未来视频帧来检测异常事件.这些方法致力于设计良好的重建模型,借助深度网络的编码-解码(预测)能力隐式地建模了上下文,并学习了鲁棒的正常时空模式.与这些方法不同,本文通过构建一个显式的上下文图,设计了时空上下文图推理模块来挖掘语义上下文信息,这有助于检测以目标为中心且上下文相关的异常事件.

## 2.2 视频上下文建模

最近一些研究方法引入了上下文模型来提高异常事件检测的性能.这些方法通过对视频上下文信息进行建模,帮助区分异常事件.本文将现有的视频上下文建模方法划分为两个类别:像素级上下文建模方法和目标级上下文建模方法,分别针对不同的粒度展开深入分析与讨论.

(1) 像素级上下文建模方法.该类方法的核心在于通过挖掘视频底层像素的时空上下文的关联性,实现对异常事件的检测.早期研究通常采用手工设计的时空特征,如三维方向梯度直方图<sup>[22]</sup>和三维尺度不变特征<sup>[23]</sup>等,来建模视频中的时空上下文.近年来,深度学习技术的发展吸引了越来越多的研究者关注深度时空特征的提取.例如,一些工作<sup>[6,18]</sup>将空间卷积的概念拓展至时间维度,创新性地提出了三维卷积自编码器结构,旨在同时整合时空上下文信息进行异常事件检测.文献<sup>[24]</sup>提出了一种动态卷积核机制,根据输入数据特性有选择性地从视频中抽取具有价值的上下文信息;文献<sup>[25]</sup>则在时空卷积网络中引入了注意力机制,使模型能够聚焦于视频中目标间的时空关系.最新的进展<sup>[26]</sup>将二维图像上的自注意力 Transformer 扩展至三维视频场景,该方法能更有效地捕获并理解不同上下文之间的复杂依赖关系,进而提升异常事件检测的性能.这些直接从像素数据中提取底层时空上下文特征的方法,易受到视频像素空间的噪声干扰,可能会误将噪声信息捕获为事件相关的特征.相比之下,本文提出的时空上下文图模型不仅能够从视频像素层面提取底层视觉特征,还进一步借助图推理方法挖掘与事件紧密相关的语义上下文信息,从而缩小了底层视觉特征与异常事件高层语义之间的语义鸿沟,并且减弱了像素空间的噪声影响,因此可以更为有效地建模与事件相关的上下文.

(2) 目标级上下文建模方法.该类方法旨在通过识别视频事件的目标,通过构建目标间的时空关

系上下文模型,进行异常事件检测.一些研究利用目标检测器<sup>[27-28]</sup>获取目标位置,它们着重于单目标的表现和时序关系上下文的建模,但往往对多个目标的空间关系和场景上下文的建模有所欠缺.例如,文献<sup>[29]</sup>采用目标检测器和光流提取器,捕获目标的表征及时序运动关系上下文,并设计了一种双流融合算法,在特征层和决策层实现上下文信息的双重整合,进行异常事件检测.与此类方法相比,本文方法则将视频事件转化为时空上下文图表示,不仅关注单目标的表现和时序运动关系的上下文,还关注不同目标间的空间关系以及所在场景的上下文建模.此外,也有一些研究关注目标时空关系的上下文建模,如文献<sup>[30]</sup>运用目标检测算法提取单个目标的表现和多个目标的空间关系,并获取骨骼特征以刻画目标的时间运动关系.作者还利用卷积神经网络独立地得到这些上下文的深层语义信息,通过简单拼接或累加方式进行异常预测.然而,这些方法通常独立且全面地提取了目标上下文,有很大可能建模了与异常事件无关的上下文信息,这会导致最终异常预测的性能降低.而本文所提出的上下文图推理方法,通过充分的信息交换与整合,有效缩小底层视觉上下文特征到异常事件语义之间的语义鸿沟.并且通过端到端的学习框架,本文方法引导模型关注与异常事件相关的上下文内容.

## 3 方 法

本文的异常检测方法有三个组成模块:上下文图表示、上下文图推理和异常预测,方法框架如图 2 所示.上下文图表示模块使用预训练的区域备选网络(Region Proposal Network, RPN)得到底层的视觉上下文特征,并建立时空上下文图.上下文图推理模块通过多个循环神经网络(Recurrent Neural Network, RNN)迭代更新上下文图的表示,以挖掘语义上下文信息.语义上下文是指在全局范围内围绕视频事件的事件目标、关系和场景.异常预测模块使用语义上下文信息来计算异常事件的分类概率.本文引入了一个自训练的方法策略,以端到端的方式联合进行上下文图推理和异常预测,从而使它们二者相互加强.一方面,上下文图推理能够提供足够的语义上下文信息以帮助检测视频异常事件;另一方面,异常预测能够指导上下文图推理,帮助其学习事件相关的语义上下文信息.

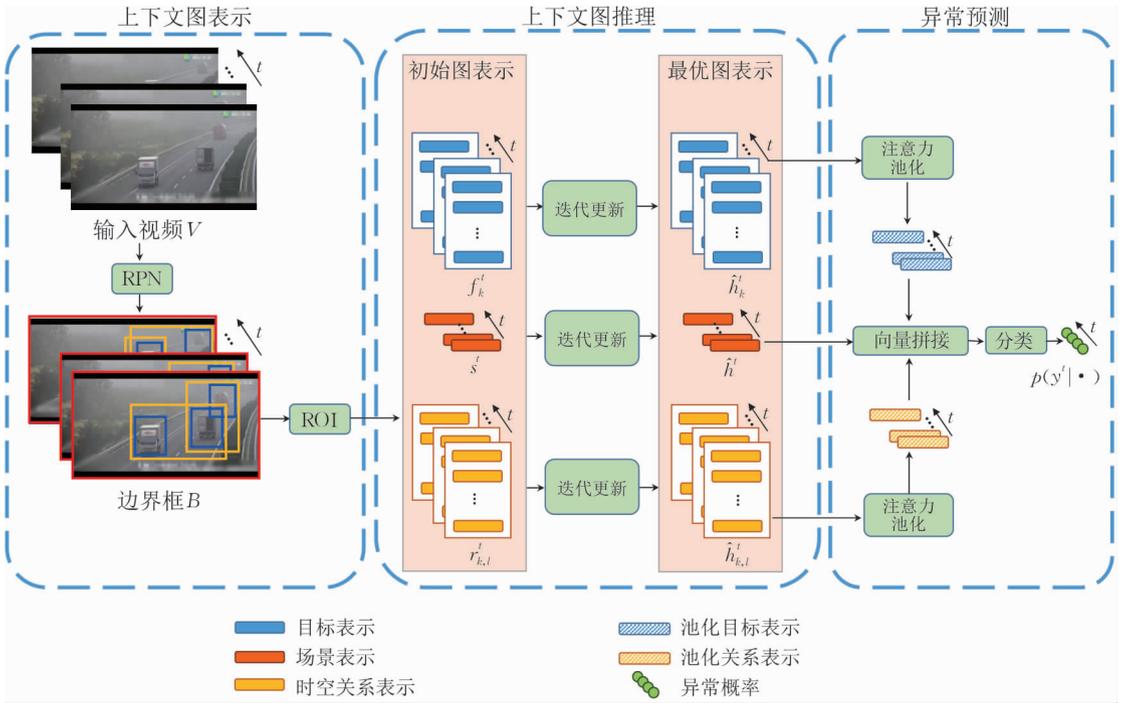


图 2 本文方法框架图

### 3.1 上下文图表示

上下文图表示旨在提取视频中事件涉及的目标、关系和场景的底层视觉特征. 输入  $T$  帧的视频  $V$ , 使用预训练的 RPN 为每个视频帧生成目标边界框, 并在第  $t$  帧中选择前  $K$  个边界框集合  $B^t = \{b_k^t | k = 1, 2, \dots, K\}$ . 同时将整个图像区域作为场景边界框. 根据目标和场景边界框, 本文方法从感兴趣区域 (Region of Interest, ROI) 网络层提取上下文表示的视觉特征, 包括目标表示、时空关系表示和场景表示; 然后建立时空上下文图, 其中图的节点表示目标/场景, 图的边表示时空关系. (1) 目标表示. 从第  $t$  帧中第  $k$  个目标的边界框  $b_k^t$  中提取视觉特征  $f_k^t$ ; (2) 时空关系表示. 先从第  $t$  帧中第  $k$  个目标和第  $l$  个目标的联合边界框提取空间关系表示的视觉特征  $r_{k,l}^t$  ( $k \neq l$ ). 然后将相邻帧的视觉特征  $f_k^t$  和  $f_k^{t+1}$  拼接为  $[f_k^t, f_k^{t+1}]$ , 该特征包含了第  $k$  个目标的时间信息. 接下来使用一个可学习矩阵  $W_d$  将拼接后的特征维度减半, 从而在第  $t$  和  $t+1$  帧生成时间关系特征  $r_{k,k}^t = W_d[f_k^t, f_k^{t+1}]$ . 为了简单的表示, 使用符号  $r_{k,l}^t$  来统一表示空间关系 ( $k \neq l$ ) 和时间关系 ( $k = l$ ); (3) 场景表示. 从  $t$  帧的场景边界框中提取视觉特征  $s^t$ .

本文以目标表示和时空关系表示为例, 介绍上下文图中上下文信息链的形成过程. 如图 3 所示, 在第  $t$  帧和  $t+1$  帧中, 第  $k, l$  个目标的空间关系表示分别为  $r_{k,l}^t$  和  $r_{k,l}^{t+1}$  ( $k \neq l$ ), 在不同帧中每个目标仅与

另一帧的该目标建立时间关系  $r_{k,k}^t$  和  $r_{k,k}^{t+1}$ . 从图中可知, 空间关系表示  $r_{k,l}^t$  和  $r_{k,l}^{t+1}$  并不直接连接形成信息链, 而是通过目标表示  $f_k^t, f_l^t, f_k^{t+1}$  和  $f_l^{t+1}$  间接形成上下文信息链. 本文的消息传递模块迭代更新上下文图中的目标表示和时空关系表示, 从而使得  $r_{k,l}^t$  和  $r_{k,l}^{t+1}$  可以进行信息交换, 即间接形成上下文信息链. 以图 3 中箭头为例, 当迭代更新次数  $n=1$  时, 相邻的第  $t$  帧和第  $t+1$  帧中的空间关系表示  $r_{k,l}^t$  和  $r_{k,l}^{t+1}$  分别将自身的信息传递并汇聚到目标表示  $(f_k^t, f_l^t)$  和  $(f_k^{t+1}, f_l^{t+1})$  上; 当  $n=2$  时,  $(f_k^t, f_l^t)$  和  $(f_k^{t+1}, f_l^{t+1})$  将汇聚后的信息传递给时间关系表示  $r_{k,k}^t$  和  $r_{k,k}^{t+1}$ , 从而形成了相邻帧之间的上下文信息链; 当  $n=3$  时, 间隔为 1 的第  $t$  帧和第  $t+2$  帧之间的空间关系表示分别为  $r_{k,l}^t$  和  $r_{k,l}^{t+2}$ , 它们借助时间关系表示  $r_{k,k}^t$  和  $r_{k,k}^{t+1}$  将信息汇聚到节点  $f_k^{t+1}$  上. 以此类推, 经过多轮迭代后, 可以形成所有帧之间的全局上下文信息链. 此外, 考虑到信息传播动力学过程的迭代往复特点, 已经交换过信息的时空关系表示会将携带的信息回传给相邻的目标表示中 (形成回路), 例如在  $n=3$  时, 空间关系表示  $r_{k,k}^t$  会将信息回传给  $f_k^{t+1}$ . 形成回路的信息交换过程可以保证在经过充分的迭代后, 每一个时空关系表示和目标表示都捕获了视频帧的全局上下文信息, 这会对检测上下文相关的异常事件提供有力的支撑和帮助.

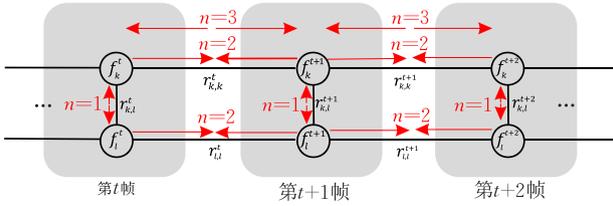


图3 目标表示和时空关系表示的连接关系示意图  
(箭头指示三轮迭代更新时信息的流动)

### 3.2 上下文图推理

上下文图推理可以从目标  $f_k^t$ 、关系  $r_{k,l}^t$  和场景  $s^t$  的视觉特征中学习事件相关的语义上下文,并缩小底层视觉特征与异常事件真实含义之间的语义差距.本文方法通过迭代更新上下文图的节点与边的表示进行上下文图推理,并通过逐步聚合各种上下文信息,捕获事件相关的语义上下文.在进行上下文图推理前,本文将目标和场景的视觉特征 ( $f_k^t, s^t$ ) 以及时空关系的视觉特征  $r_{k,l}^t$  分别初始化为图节点和边的初始表示,它们的集合为  $c = \{f_k^t, r_{k,l}^t, s^t | t=1, 2, \dots, T; k, l=1, 2, \dots, K\}$ ; 在上下文图推理过程中,根据上下文图的依赖关系对这些图表示进行迭代更新,得到图表示的集合为  $h = \{h_k^t, h_{k,l}^t, h^t | t=1, 2, \dots, T; k, l=1, 2, \dots, K\}$ ; 最终目标是获得具有语义上下文信息的最优图表示  $\hat{h}$ . 本文获取  $\hat{h}$  的目标公式为

$$\hat{h} = \arg \max_h P(h|c) \quad (1)$$

已有的文献[31]表明,条件随机场(Conditional Random Fields, CRF)可以有效地建模图模型.因此,式(1)可以建模为一个有着上下文图结构的条件随机场过程,其中后验边际分布  $P(h|c)$  视为一个吉布斯分布(Gibbs distribution),即  $P(h|c) = 1/Z(c) \exp(-E(h|c))$ , 这里  $E(h|c)$  是将  $\hat{h}$  分配给  $h$  的吉布斯能量(energy),  $1/Z(c)$  表示归一化的配分函数(partition function). 为了简化符号,后文公式中省略了条件  $c$ , 并且用字符  $h_i$  和  $\hat{h}_i$  分别表示集合  $h$  和  $\hat{h}$  中的第  $i$  个元素. 条件随机场的吉布斯能量为

$$E(h) = \sum_i \varphi_u(h_i) + \sum_{(i,j) \in o} \varphi_p(h_i, h_j) \quad (2)$$

其中  $\varphi_u(h_i)$  表示一元能量,度量了  $h_i$  取得最优表示  $\hat{h}_i$  的逆似然(即成本);  $\varphi_p(h_i, h_j)$  表示二元能量,度量了  $h_i$  和  $h_j$  同时取得最优表示  $\hat{h}_i$  和  $\hat{h}_j$  的逆似然; 集合  $o$  表示图中相互连通元素的索引  $(i, j)$  的集合. 在本文的上下文图中,目标节点以及与该节点有关的时空关系边是相互连通的,场景节点与所有目标节点和时空关系边是相互连通的,这些连通的节点与边的索引构成了集合  $o$ . 一元能量  $\varphi_u(h_i)$  不需要

考虑其他上下文的影响,可以用简单的独立分类器来求解,而二元能量  $\varphi_p(h_i, h_j)$  则需要考虑其他上下文的影响.

最小化能量  $E(h)$ , 等价于最大化将  $\hat{h}$  分配给图表示的可能性. 由于在实际中难以直接计算这种最小化问题的解析解,因此可以使用平均场(mean-fields)理论<sup>[32]</sup> 近似求解最大后验分布  $P(h)$ . 平均场近似的目标是求解满足最小化 KL 散度(Kullback-Leibler divergence)  $KL(P(h) \parallel Q(h))$  的更简单的近似分布  $Q(h)$ , 并且  $Q(h)$  满足  $Q(h) = \prod_i Q_i(h_i)$ . 分布  $Q(h)$  的迭代公式为

$$Q_i(h_i) = \frac{1}{Z_i} \exp(-\varphi_u(h_i) - \sum_{(i,j) \in o} \mathbf{E}_{U_j \sim Q_j} [\varphi_p(h_i, U_j)]) \quad (3)$$

式(3)的详细推导过程见附录A. 在传统的求解方法中,式(3)中左右两边都有分布  $Q(h)$ , 因此往往先初始化一个  $Q(h)$ , 然后不断迭代代入式(3)中更新,直到  $P(h)$  和  $Q(h)$  的 KL 散度小于一定阈值(附录A给出了常用的迭代算法). 然而,这种迭代更新方法并不适用于本文,原因是本文并没有最优图表示集合  $\hat{h}$  的真值标注,无法直接计算能量函数  $\varphi_u(h_i)$  和  $\varphi_p(h_i, h_j)$ . 受益于深度学习的发展,本文可以通过神经网络的前向过程,将当前迭代得到图表示传递给后面的异常预测模块; 然后根据预测的误差衡量当前图表示的优劣,以此计算梯度并通过梯度的反向传播来优化图的上下文表示; 最终可以得到使得异常预测误差最小的  $\hat{h}$  (即与异常事件语义最相关的  $\hat{h}$ ). 已有的文献[33]证明了通过使用消息传递模块的深度循环神经网络(Recurrent Neural Network, RNN), 可以实现平均场近似过程中分布  $Q(h|c)$  的建模. 受到这些工作的启发,本文将满足独立边际分布乘积形式的  $Q(h|c)$  写为

$$Q(h|c) = \prod_{t=1}^T Q(\hat{h}^t | h^t) Q(h^t | s^t) \prod_{k=1}^K Q(\hat{h}_k^t | h_k^t) Q(h_k^t | f_k^t) \prod_{l=1}^K Q(\hat{h}_{k,l}^t | h_{k,l}^t) Q(h_{k,l}^t | r_{k,l}^t) \quad (4)$$

本文使用3个独立的循环神经网络,包括  $RNN_{obj}$ 、 $RNN_{rel}$  和  $RNN_{sce}$ , 来计算和迭代更新目标表示、时空关系表示和场景表示在  $Q(h|c)$  中的状态,如图4所示. 在迭代更新过程中,本文对所有的上下文表示进行建模,充分探索了上下文信息. 在  $RNN_{obj}$  的第  $n$  次迭代中,  $h_k^{t,n}$  是第  $k$  个目标的图表示,  $h_{k,l}^{t,n}$  和  $h_{l,k}^{t,n}$  是第  $k$  个目标和第  $l$  个目标之间的关系表示,  $h^{t,n}$  是场景表示.

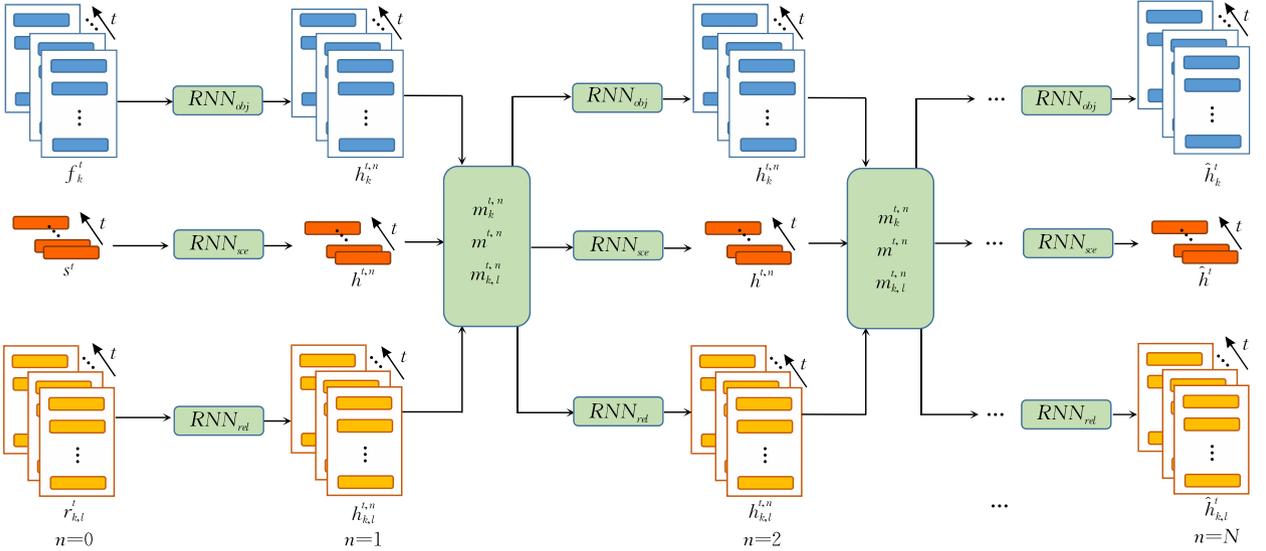


图 4 迭代更新过程示意图

目标表示  $h_k^{t,n+1}$  迭代到第  $n+1$  时刻的计算公式表示为

$$\begin{aligned} h_k^{t,n+1} &= RNN_{obj}(m_k^{t,n}, h_k^{t,n}), \\ m_k^{t,n} &= \sum_l \sigma(\mathbf{W}_f^{1\top} [h_k^{t,n}, h_{k,l}^{t,n}]) h_{k,l}^{t,n} + \\ &\quad \sum_l \sigma(\mathbf{W}_f^{2\top} [h_k^{t,n}, h_{l,k}^{t,n}]) h_{l,k}^{t,n} + \\ &\quad \sigma(\mathbf{W}_f^3 [h_k^{t,n}, h^{t,n}]) h^{t,n} \end{aligned} \quad (5)$$

其中  $\sigma$  表示 Sigmoid 激活函数,  $[\cdot, \cdot, \cdot, \cdot]$  表示向量拼接.  $\mathbf{W}_f^1, \mathbf{W}_f^2$  和  $\mathbf{W}_f^3$  是可学习的网络参数.  $RNN_{obj}$  表示循环网络函数. 类似地, 时空关系表示  $h_{k,l}^{t,n+1}$  的更新过程为

$$\begin{aligned} h_{k,l}^{t,n+1} &= RNN_{rel}(m_{k,l}^{t,n}, h_{k,l}^{t,n}), \\ m_{k,l}^{t,n} &= \sigma(\mathbf{W}_r^{1\top} [h_{k,l}^{t,n}, h_k^{t,n}]) h_k^{t,n} + \\ &\quad \sigma(\mathbf{W}_r^{2\top} [h_{k,l}^{t,n}, h_l^{t,n}]) h_l^{t,n} + \\ &\quad \sigma(\mathbf{W}_r^3 [h_{k,l}^{t,n}, h^{t,n}]) h^{t,n} \end{aligned} \quad (6)$$

其中  $\mathbf{W}_r^1, \mathbf{W}_r^2$  和  $\mathbf{W}_r^3$  是可学习的参数,  $RNN_{rel}$  表示循环网络函数. 场景表示  $h_{k,l}^{t,n+1}$  的更新过程为

$$\begin{aligned} h_{k,l}^{t,n+1} &= RNN_{sce}(m_{k,l}^{t,n}, h_{k,l}^{t,n}), \\ m_{k,l}^{t,n} &= \sum_k \sigma(\mathbf{W}_s^{1\top} [h_{k,l}^{t,n}, h_k^{t,n}]) h_k^{t,n} + \\ &\quad \sum_{k,l} \sigma(\mathbf{W}_s^2 [h_{k,l}^{t,n}, h_{k,l}^{t,n}]) h_{k,l}^{t,n} \end{aligned} \quad (7)$$

$\mathbf{W}_s^1$  和  $\mathbf{W}_s^2$  是可学习的参数,  $RNN_{sce}$  表示循环函数.

### 3.3 异常预测

异常预测旨在基于推理的语义上下文预测第  $t$  帧视频的标签  $y^t$  ( $y^t = 1$  表示预测为异常,  $y^t = 0$  表示正常). 本文通过最大化后验概率分布来求解异常预测问题, 即

$$P(y^t | \{\hat{h}_k^t, \hat{h}_{k,l}^t, \hat{h}^t | k, l = 1, 2, \dots, K\}) \quad (8)$$

异常事件通常涉及不同的上下文信息, 这意味

着上下文表示对区分各类异常事件的贡献是不同的. 因此, 衡量不同上下文表示的贡献有助于异常检测. 为此, 本文使用两个独立的注意力池化 (attention-pooling) 网络对所有目标表示  $\hat{h}_k^t$  和所有关系表示  $\hat{h}_{k,l}^t$  进行编码. 对于第  $t$  帧, 注意力池化的表示  $f^t$  由式 (9) 给出:

$$f^t = \sum_{k=1}^K \hat{h}_k^t \cdot \alpha_k^t, \quad \alpha_k^t = \frac{\exp(\mathbf{W}_p \hat{h}_k^t)}{\sum_{l=1}^K \exp(\mathbf{W}_p \hat{h}_l^t)} \quad (9)$$

其中  $\mathbf{W}_p$  是注意力池化网络的参数. 池化的关系表示  $r^t$  是通过另一个注意力池化网络以类似的方式计算得到. 然后将池化目标表示  $f^t$ 、池化关系表示  $r^t$  和场景表示  $\hat{h}^t$  拼接起来, 计算异常的概率分布, 即

$$P(y^t | \cdot) = \phi(\mathbf{W}_o \sigma(\mathbf{W}_y [f^t, r^t, \hat{h}^t])) \quad (10)$$

其中  $\phi$  是计算概率向量的 SoftMax 函数,  $\mathbf{W}_o$  和  $\mathbf{W}_y$  是可学习的参数.

### 3.4 学习策略

在上下文图推理模块和异常预测模块中, 所有可学习的网络参数都以端到端的方式, 通过反向传播的随机梯度下降 (Stochastic Gradient Descent, SGD) 算法进行优化. 为此, 本文分别在无监督、半监督、弱监督和监督学习方式下设计了相应的损失函数.

(1) 无监督学习. 在无监督学习方式下, 正常和异常数据的标注均无法获取. 为了训练网络, 本文引入了一种简单的自训练 (self-training) 策略: 先获取数据的伪标签, 然后对它们进行迭代优化. 具体来说, 给定一个包含  $T$  帧的视频  $V$ , 本文使用预训练的 RPN 生成目标和场景的边界框, 并提取目标表示  $f_k^t$  和场景表示  $s^t$ . 然后聚合这些表示以生成帧级表示  $[\sum_k f_k^t / K, s^t]$ . 本文通过引入无监督扩展的孤

立森林(Extended Isolation Forest, EIF)方法<sup>[34]</sup>初始化伪标签. EIF 以  $[\sum_k f'_k/K, s^t]$  为输入, 并对视频帧进行异常分类. 本文对 EIF 的路径长度进行倒叙排序, 并提取前  $M_1$  个和后  $M_2$  个样本分别构造正常数据集  $\mathcal{P}$  和异常数据集  $\mathcal{N}$ , 这样就得到了伪标注的训练数据. 为了训练网络, 本文从  $\mathcal{P}$  和  $\mathcal{N}$  中选择三元组样本  $\{V_a, V_p, V_n\}$ , 其中  $V_a$  表示锚样本(正常事件),  $V_p$  是正样本(正常事件),  $V_n$  表示负样本(异常事件). 将样本输入到本文的异常检测方法中, 并通过式(10)计算它们的异常概率  $P'_a, P'_p$  和  $P'_n$ , 然后使用损失函数  $\mathcal{L}_1$  来训练网络, 即

$$\mathcal{L}_1 = \mathcal{L}_1^{cls} + \lambda \mathcal{L}_1^{tri} \quad (11)$$

其中,  $\mathcal{L}_1^{cls}$  是单类分类器的损失函数,  $\mathcal{L}_1^{tri}$  是三元组损失函数,  $\lambda$  是权衡参数.  $\mathcal{L}_1^{cls}$  和  $\mathcal{L}_1^{tri}$  的公式为

$$\begin{aligned} \mathcal{L}_1^{cls} &= \frac{1}{T} \sum_{t=1}^T \log(1 - P'_a), \\ \mathcal{L}_1^{tri} &= \frac{1}{T} \sum_{t=1}^T \max(0, \|P'_a - P'_n\|_2 - \\ &\quad \|P'_a - P'_p\|_2 + \delta) \end{aligned} \quad (12)$$

$\delta$  表示三元组损失函数的间隔参数. 本文使用单类分类器  $\mathcal{L}_1^{cls}$  分类正常视频, 而没有选择使用二分类器同时对正常视频和异常视频进行分类. 原因如下: 正常视频的模式可以被较好地概括总结, 因此可以使用单类分类器进行建模; 而异常视频有更复杂的模式, 难以被分为一个单一的类别, 这对应着异常事件的无边界特性(unbounded nature). 因此, 本文引入了三元组损失函数  $\mathcal{L}_1^{tri}$ , 目的是减少正常视频类内差异, 并增加正常视频和异常视频之间的类间差异. 本文使用损失函数  $\mathcal{L}_1$  训练网络后, 再使用式(10)计算异常概率  $P'_a, P'_p$  和  $P'_n$ , 并再次选择前  $M_1$  个样本和后  $M_2$  个本来调整标注集合  $\mathcal{P}$  和  $\mathcal{N}$ . 调整后标注数据可以再次输入到式(11)中微调网络. 本文通过迭代地调整伪标注数据和微调网络, 实现稳定的检测性能.

(2) 半监督学习. 在半监督学习方式中, 只有正常数据可于网络训练. 与无监督学习类似, 本文获取表示  $[\sum_k f'_k/K, s^t]$ , 然后使用 EIF 生成异常数据的初始化伪标签集  $\mathcal{N}$ , 而正常数据的标签集  $\mathcal{P}$  是已知的. 本文将这些标注的数据  $\mathcal{P}$  和  $\mathcal{N}$  输入到网络中, 并使用式(11)中的损失函数  $\mathcal{L}_1$  来迭代进行异常预测和调整集合  $\mathcal{N}$ .

(3) 弱监督学习. 在弱监督学习方式下, 只有正常数据和带有视频级标注的异常数据可用于训练网络. 在异常预测中, 只有视频级的标注可以使用, 但并不知道具体哪帧是正常或异常的. 在该情况下, 本

文首先通过半监督学习方法, 使用所有正常数据对网络进行预训练. 然后使用预训练模型计算异常视频中每个帧的异常概率分布. 考虑到每个异常视频至少包含一个异常帧, 可以将视频中所有帧的异常概率归一化到  $[0, 1]$ , 即

$$\gamma^t = \frac{P^t - \min_t P^t}{\max_t P^t - \min_t P^t} \quad (13)$$

其中  $\gamma^t$  是预训练模型的归一化异常分数. 本文将异常概率大于 0.5 的帧标记为异常视频帧. 然后从正常视频帧和异常视频帧中选择三元组, 并通过优化损失函数  $\mathcal{L}_2$  来训练网络, 即

$$\mathcal{L}_2 = \mathcal{L}_2^{cls} + \lambda \mathcal{L}_2^{tri} \quad (14)$$

单类分类器损失函数  $\mathcal{L}_2^{cls}$  和加权三元组损失函数  $\mathcal{L}_2^{tri}$  为

$$\begin{aligned} \mathcal{L}_2^{cls} &= \frac{1}{T} \sum_{t=1}^T \log(1 - P'_a), \\ \mathcal{L}_2^{tri} &= \frac{1}{T} \sum_{t=1}^T \gamma^t \max(0, \|P'_a - P'_n\|_2 - \\ &\quad \|P'_a - P'_p\|_2 + \delta) \end{aligned} \quad (15)$$

(4) 监督学习. 在监督学习方式下, 本文使用带有帧级标注的异常数据和正常数据来训练网络. 本文从这些数据中选择三元组样本, 并使用式(11)中的损失函数  $\mathcal{L}_1$  训练网络.

### 3.5 异常评分

在测试期间, 通过一次网络的前向传递过程, 本文计算得到式(10)中第  $t$  帧的异常概率, 然后将之转换为异常分数, 并进行异常预测. 由于异常分数应在帧之间平滑变化, 本文仿照文献[21]使用高斯滤波器强制帧级异常分数保持平滑; 为了确保实验结果的可比性和公正性, 滤波器参数和文献[21]保持一致, 即高斯核的窗口半径 125, 方差 30.

## 4 实 验

### 4.1 数据集

本文在 UCF-Crime<sup>[36]</sup>、Subway<sup>[38]</sup>、Avenue<sup>[39]</sup> 和 ShanghaiTech<sup>[40]</sup> 数据集上进行了实验. UCF-Crime 数据集是一个包含视频级标注的弱监督数据集, 本文以半监督、弱监督和监督方式在该数据集进行了实验. Subway、Avenue 和 ShanghaiTech 是半监督数据集, 本文以无监督和半监督方式在这些数据集上训练了网络. 对于无监督学习, 本文按照文献[10]将训练集和测试集合并为无监督训练的数据集. 对于半监督学习, 本文使用训练集中所有正常数据进行训练, 使用测试集进行测试. (1) UCF-Crime 是一个大规模的真实世界监控视频数据集, 在不同场景

中记录了 13 类异常事件, 包括 1610 个训练视频和 290 个测试视频. 对于半监督学习, 本文选择所有正常的训练视频来训练网络. 对于弱监督学习, 本文使用训练集中带有弱监督标注的正常和异常数据训练网络. 对于监督学习, 本文从每个异常训练视频中手动标记 20 s 左右的异常帧, 这些视频帧形成一个新的异常训练集; (2) Subway 数据集包含两个场景: Entrance(144 249 个视频帧) 和 Exit(64 900 个视频帧). 异常事件包括错误方向、逃票、闲逛等. 本文遵

循文献[4]中的设置来划分训练和测试集: Entrance 和 Exit 视频的前 15 min 属于训练集, 其余构成测试集; (3) Avenue 数据集包括 16 个训练视频和 21 个测试视频, 总帧数为 35 240 帧. 异常事件包括跑步、反方向行走、投掷物品和闲逛等; (4) ShanghaiTech 数据集包含 13 个视角的场景, 具有复杂的光照条件. 该数据集有 130 个异常事件和超过 270 000 帧训练视频. 4 个数据集中典型的异常事件示例如图 5 所示.



图 5 在 UCF-Crime、Subway、Avenue 和 ShanghaiTech 数据集中典型的异常事件示例(边界框内)

## 4.2 实验细节

本文用 RPN<sup>[33]</sup> 和 KLT (Kanade-Lucas-Tomasi) 跟踪器<sup>[41]</sup> 来生成目标边界框. 本文选择前 10 (即  $K=10$ ) 个边界框, 得到相对冗余的上下文信息. KLT 跟踪器通过连接相邻帧中的每个目标, 帮助建模目标的时间关系. 本文按照文献[33]的实验设置构建 RPN 和 RNN, 并将图表示的维度设置为 512. 在所有学习方式中, 本文选择 10 帧固定长度的滑动窗口作为输入, 设置训练批大小 (batch size) 为 3, 并使用学习率为 0.001 的 RMSprop 优化器来训练网络. 本文设置间隔参数  $\delta$  为 0.4, 权衡参数  $\lambda$  为 1.0. 在初始

化伪标签时, 本文使用 PCA 从  $[\sum_k f_k^t / K, s^t]$  中提取 128 维表示, 并设置 EIF 的扩展级别为 64. 伪标签集  $\mathcal{P}$  和  $\mathcal{N}$  的大小根据经验分别设置为相应训练集大小的 1/3 和 1/10. 迭代次数设置为 2.

## 4.3 评估指标

本文计算帧级异常分数, 并分别在帧级和事件级上进行异常检测的方法评估. 本文通过逐渐改变异常分数阈值, 来绘制接受者操作特征曲线 (Receiver Operating Characteristic Curve, ROC). 然后使用相应的曲线下面积 (Area Under Curve, AUC) 进行帧级评估. 本文还引入了误报率 (False Alarm)

来评估错误分类的概率. 较高的  $AUC$  值和较低的误报率表示更好的帧级异常检测性能, 帧级比较结果见表 1 和表 2. 此外, 本文还仿照文献[4]聚合连续帧的异常分数, 得到事件级的检测结果. 本文计算了检测到的事件计数(event count), 并统计了正确检测到的异常样本数量和错误检测到的正常样本数量, 以此进行事件级异常检测的评估, 事件级比较结果见表 3. 本文除了对 Ionescu 等人<sup>[5]</sup>和 Hariri 等人<sup>[34]</sup>的工作进行了复现之外, 表 1、表 2 和表 3 内其他用于比较的方法的评估指标均取自对应论文的实验数据.

表 1 在 UCF-Crime 数据集上的帧级结果比较

训练方式	方法	$AUC/\%$	$False\ Alarm/\%$
半监督	Hasan 等人 <sup>[4]</sup>	50.6	27.2
	Ionescu 等人 <sup>[5]</sup>	66.1	8.5
	Wang 等人 <sup>[35]</sup>	70.5	2.1
	Sun 等人 <sup>[11]</sup>	72.7	2.2
	Ours	74.8	2.0
弱监督	Sultani 等人 <sup>[36]</sup>	75.4	1.9
	Sapkota 等人 <sup>[37]</sup>	83.4	N/A
	Ours	83.8	1.4
监督	SVM 基线	72.8	2.3
	MLP 基线	75.3	1.9
	Ours	84.5	1.2

表 2 在 Subway、Avenue 和 ShanghaiTech 数据集上的帧级异常检测结果比较( $AUC$ )

训练方式	方法	$AUC/\%$			
		Subway (Entrance)	Subway (Exit)	Avenue	ShanghaiTech
半监督	Hasan 等人 <sup>[4]</sup>	94.3	80.7	70.2	N/A
	Ionescu 等人 <sup>[17]</sup>	70.6	85.7	80.6	N/A
	Wang 等人 <sup>[42]</sup>	N/A	84.5	85.3	N/A
	Chong 和 Tay <sup>[3]</sup>	84.7	94.0	80.3	N/A
	Song 等人 <sup>[43]</sup>	90.2	94.6	89.2	70.0
	Sun 等人 <sup>[11]</sup>	N/A	N/A	89.6	74.7
	Feng 等人 <sup>[12]</sup>	N/A	N/A	85.9	77.7
	Cai 等人 <sup>[13]</sup>	N/A	N/A	87.4	74.2
	Yang 等人 <sup>[26]</sup>	N/A	N/A	89.9	73.8
	Wu 等人 <sup>[24]</sup>	N/A	N/A	90.6	75.5
	Wang 等人 <sup>[25]</sup>	N/A	N/A	86.1	73.2
	Doshi 等人 <sup>[30]</sup>	N/A	N/A	85.8	71.2
	Yang 等人 <sup>[29]</sup>	N/A	N/A	84.2	83.8
	Ours	91.4	95.2	91.4	78.5
无监督	Hasan 等人 <sup>[4]</sup>	68.8	75.9	70.1	62.7
	Hariri 等人 <sup>[34]</sup>	83.8	92.7	78.9	70.3
	Yu 等人 <sup>[44]</sup>	N/A	N/A	90.7	72.6
	Ours	90.1	93.0	90.0	73.2

表 3 在 Subway 和 Avenue 数据集上的事件级异常检测结果比较(异常检测数/异常误报数)

方法	异常检测数/异常误报数		
	Subway (Entrance) GT: 66	Subway (Exit) GT: 19	Avenue GT: 47
Lu 等人 <sup>[39]</sup>	57/4	19/2	N/A
Kim 等人 <sup>[47]</sup>	56/3	19.2	N/A
Dutta 等人 <sup>[48]</sup>	60/5	19/2	N/A
Hasan 等人 <sup>[4]</sup>	61/15	17/5	45/4
Chong 和 Tay <sup>[3]</sup>	61/9	18/10	44/12
Ours	61/7	19/2	46/3

#### 4.4 实验比较

(1) UCF-Crime 数据集的结果. 表 1 展示了在半监督、弱监督和监督学习方式下, 在 UCF-Crime 数据集上的  $AUC$  性能和误报率的比较. 除了 Ionescu 等人<sup>[5]</sup>的工作外, 其他比较方法的性能指标均取自原始文献. 本文复现了 Ionescu 等人的工作, 将他们使用的目标检测器替换为本文使用的 RPN, 目的是为了进行公平比较.

在半监督学习方式下, 与现有方法<sup>[11,35]</sup>相比本

文方法实现了最先进的性能,  $AUC$  指标分别提高了 4.3% 和 2.1%, 这验证了本文方法在各种场景中检测上下文相关异常事件的优越性. 弱监督学习方式下, 本文方法的  $AUC$  与弱监督方法<sup>[37]</sup> 相比提高了 0.4%, 这表明本文方法可以有效地检测未知异常事件.

在监督学习方式下, 本文方法与 SVM 基线和 MLP 基线方法进行了比较, 在  $AUC$  评估上实现了 11.7% 和 9.2% 的性能提升. SVM 和 MLP 方法均以视频帧中的初始目标表示、关系表示和场景表示为输入. MLP 的结构为 FC(512, 256, ReLU)-FC(256, 128, ReLU)-FC(128, 2, SoftMax), 其中  $FC(a, b, f)$  表示具有可训练权重矩阵为  $\mathbf{W} \in \mathbb{R}^{a \times b}$  和激活函数为  $f$  的全连接层. SVM 和 MLP 在训练过程中将标记为异常帧的所有表示分为一类, 将正常帧的所有表示分为另一类; 在测试过程中将每一帧中所有上下文表示的最高异常类分数作为帧级异常分数.

图 6 展示了 ROC 曲线以进行进一步的评估. 本

文方法在弱监督和监督学习方式中的曲线几乎完全包裹现有方法的曲线, 这意味着本文方法在各种阈值下优于已有工作<sup>[4,5,36,39]</sup>. 从图中还可观察到, 本文方法在半监督学习方式下的性能甚至可以与弱监督方法<sup>[36]</sup>相媲美.

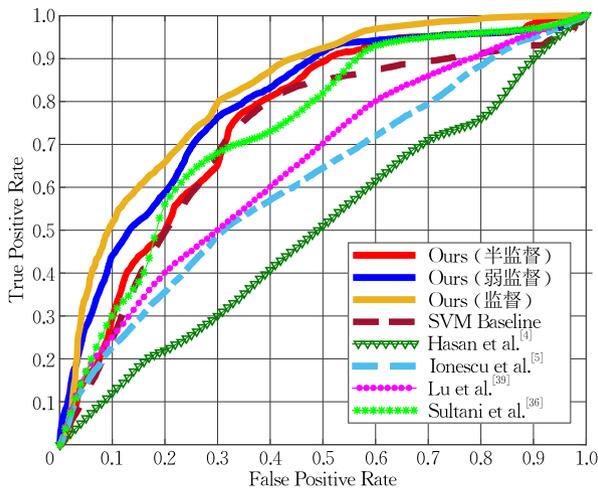


图 6 ROC 曲线图

(2) Subway 数据集的结果. 本文在表 2 中报告了 Subway 数据集的结果. 在半监督学习方式中, 本文方法比大多数比较方法效果要好. 在 Entrance 和 Exit 数据集上 AUC 的性能提升验证了建模上下文对于区分异常事件的重要性. 在无监督学习方式下, 本文方法也取得了可比较的性能, 在 Exit 数据集上获得了最高的 AUC 指标, 达到了 95.2%, 在 Entrance 数据集上获得了可比较的 91.4% AUC 值. 从 Subway 数据集的整体性能来看, 本文方法与最先进的方法相比取得了可比较的结果.

(3) Avenue 数据集的结果. 表 2 展示了无论是在半监督学习还是无监督学习方式下, 本文方法在 Avenue 数据集上均取得了有竞争力的实验结果. 对于半监督学习, Doshi 等人<sup>[30]</sup>通过运用 Yolo-v5<sup>[45]</sup>这一优秀的目标检测器实现了高至 85.8% 的优异 AUC 性能; 而 Yang 等人<sup>[26]</sup>利用了 Transformer<sup>[46]</sup>等先进网络结构进一步将 AUC 提升至 89.9%. 本文使用上下文建模和推理的方法, 成功地获得了最高的 91.4% 的 AUC 指标. 此外, 相较于本文方法的早期版本<sup>[11]</sup>, 本文的改进方法在 AUC 评估上获得了 1.8% 的提升. 在 Avenue 数据集上, 对于无监督学习, 本文方法与最先进的帧级异常检测方法<sup>[44]</sup>相比, 也取得了可比较的结果. 表 2 的结果验证了本文方法是有效和稳健的.

(4) ShanghaiTech 数据集的实验结果. 表 2 展示了本文方法在 ShanghaiTech 数据集上的结果.

ShanghaiTech 数据集场景复杂、动作多样, 具有很大的挑战性. 在半监督学习方式下, 在与多种对比方法的比较中, 本文方法取得了次优的 AUC 性能, 仅次于 Yang 等人<sup>[29]</sup>的方法. 这一差异可归因于 Yang 等人<sup>[29]</sup>使用了 RGB 图像、光流图像和骨骼特征等多种输入信息, 而本文仅以 RGB 图像作为单一输入源. 尽管如此, 根据 Yang 等人<sup>[29]</sup>在其论文中的消融实验表明, 当去除光流图像和骨骼特征后其方法的 AUC 值降至 70.8%, 而我们的方法 AUC 的结果要更高. 此外, 相较本文方法的早期版本<sup>[11]</sup>, 本文 AUC 提高了 3.8%. 在无监督学习方式下, 本文方法比 Yu 等人<sup>[44]</sup>的方法 AUC 提高了 0.6%. 综上所述, 在 ShanghaiTech 数据集的实验结果能够验证本文提出的上下文建模和推理方法在异常事件检测任务上的有效性.

#### 4.5 事件计数

本文设置异常分数阈值来检测异常事件, 并计算事件计数进行评估, 结果见表 3. 按照文献<sup>[4]</sup>中的设置, 本文假设 50 帧内的局部异常分数的极值属于同一异常事件, 目的是减少异常分数中的噪声. 表 3 展示了两个数据集上检测到的异常事件和误报的数量. 对于 Avenue 数据集, 本文方法可以比工作<sup>[3-4]</sup>更准确地检测到异常事件. 对于 Subway 数据集, 与最先进的事件级异常检测方法<sup>[48]</sup>相比, 本文方法实现了可比较的性能. 结果表明, 本文方法可以更准确地定位异常事件的时间区域, 在实际场景中更实用.

#### 4.6 消融实验

本文在 Subway、Avenue 和 ShanghaiTech 三个数据集上, 以半监督和无监督的学习方式比较了本文方法的不同模块的贡献, 消融实验结果见表 4. “w/o 空间关系”表示移除空间关系的建模(剔除掉上下文图的空间的边), 但是保留时间关系的建模. “w/o 时间关系”表示剔除掉上下文图时间的边, 仅在单个帧中对目标、空间关系和场景的表示进行上下文图推理. “w/o 时空关系”是指忽略时空关系(剔除掉上下文图的时空的边), 仅仅根据目标和场景的上下文信息来区分异常. “w/o 场景”是指移除场景上下文的建模. “w/o 上下文图推理”是指移除迭代更新的上下文图推理模块, 直接将初始化的图表示, 即目标、关系和场景表示, 输入到异常预测模块中. “w/o 注意力池化”是在异常预测模块中使用简单的平均池化操作取代注意力池化网络层, 进行上下文信息聚合.

表 4 本文方法在 Subway、Avenue 和 ShanghaiTech 数据集上的消融实验结果

训练方式	方法	AUC/%			
		Subway (Entrance)	Subway (Exit)	Avenue	ShanghaiTech
半监督	w/o空间关系	87.4	90.1	88.7	75.2
	w/o时间关系	90.7	94.6	89.8	75.9
	w/o时空关系	85.1	86.3	85.0	74.5
	w/o场景	91.0	94.5	90.9	76.8
	w/o上下文图推理	87.9	91.9	87.4	74.2
	w/o注意力池化	84.4	86.2	86.6	74.0
	Ours	91.4	95.2	91.4	78.5
无监督	w/o空间关系	86.6	90.2	88.8	69.9
	w/o时间关系	88.9	91.0	89.4	71.3
	w/o时空关系	84.0	86.7	85.3	68.4
	w/o场景	89.5	92.0	89.5	72.2
	w/o上下文图推理	87.4	89.9	88.0	70.3
	w/o注意力池化	85.6	88.1	87.4	70.0
	Ours	90.1	93.0	90.0	73.2

从表 4 中观察到:(1) 根据“w/o空间关系”、“w/o时间关系”和“w/o时空关系”的实验结果,得出无论是在半监督学习还是无监督学习中,当消融方法不考虑空间或时间关系时,AUC 的性能都显著降低,这验证了时空关系上下文对于区分异常事件的重要性.此外,“w/o场景”的实验结果表明考虑场景上下文能显著提高性能,证明了识别场景类型对检测异常事件是很有帮助的;(2) 根据“w/o上下文图推理”的实验结果,当消融方法不进行上下文图推理时,在半监督训练中,AUC 指标在 Subway(Entrance/Exit)、Avenue 和 ShanghaiTech 数据集上分别下降了 3.5%/3.3%/4.0%和 4.3%;在无监督训练中,AUC 指标在三个数据集上分别降低了 2.7%/3.1%、2.0%和 2.9%.这验证了上下文图推理在视频异常事件检测中的有效性;(3) “w/o注意力池化”的结果表明,本文使用的注意力池化网络也有助于提升 AUC 的性能,该池化网络是一种有效的上下文信息聚合方法.

#### 4.7 定性结果

图 7 展示了 UCF-Crime 和 ShanghaiTech 数据集上检测到的异常事件的三个示例图,其中深色窗口表示异常的真值标签,曲线表示测试视频中部分帧的异常分数.图中检测到的异常事件是 UCF-Crime 数据集上的“抢劫”和“入店行窃”以及 ShanghaiTech 数据集上的“奔跑”和“跳跃”.从图 7 中可以清楚地看到本文方法所产生的异常分数与其真值标签匹配得很好.并且正常样本和异常样本之间存在较大的分数差距,验证了本文方法的有效性.

为了进一步验证上下文建模与推理在视频异常事件检测中的作用,本文在图 8 中展示了在半监督学习下的 Avenue 数据集上“帧级异常分数-帧级特征”

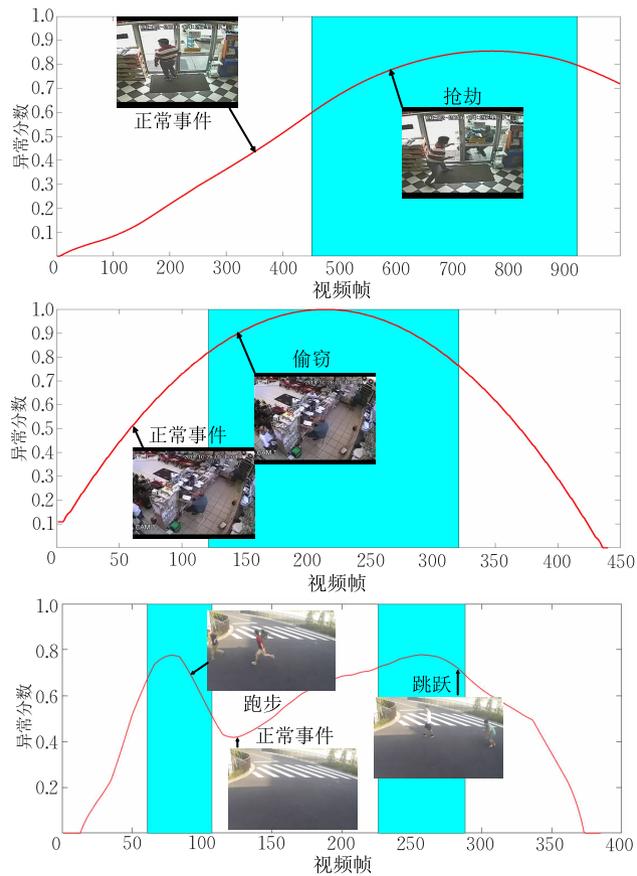


图 7 视频帧异常分数曲线图

关系的两个可视化实例,证明了本文方法得到的异常分数具有更强的区分性.具体而言,图 8(a)展示了采用完整模型(Ours)的结果;图 8(b)是未采用上下文图推理(w/o上下文图推理)的消融实验版本的结果.其中,本文方法所使用的视频帧特征是式(9)中帧级注意力池化的表示  $f^t$ ;而在 w/o上下文图推理方法中,通过用图的初始表示  $f_k^i, r_{k,i}^i$  和  $s^i$  替换掉式(9)中的输入  $\hat{h}_k^i, \hat{h}_{k,i}^i$  和  $\hat{h}^i$  (即不进行图推理),得到帧级特

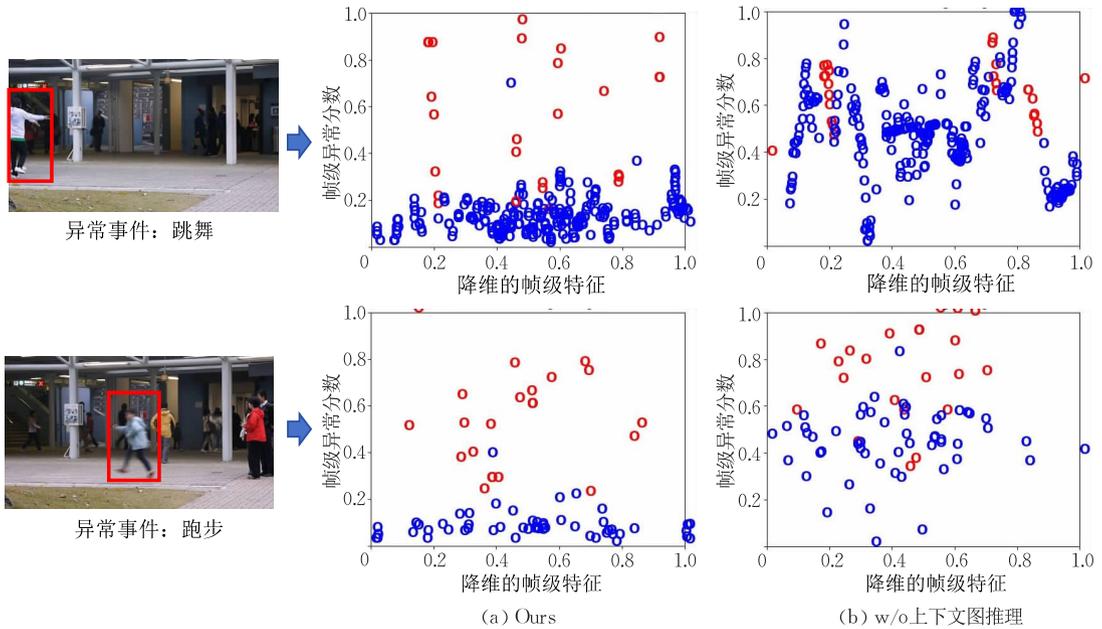


图 8 半监督学习下在 Avenue 数据集上“帧级异常分数-帧级特征”关系的两个示例图

征并计算异常分数. 图 8 中红色和蓝色的点分别表示视频帧中异常帧和正常帧, 坐标横轴表示使用 t-SNE 算法对视频帧特征降维后得到的一维特征 (min-max 归一化), 纵轴表示计算得到的异常分数. 可以观察到本文方法的帧级异常分数具有更好的区分性.

#### 4.8 检测时间分析

表 5 展示了在 Subway、Avenue 和 ShanghaiTech 三个数据集上的测试时间, 本文在测量时并未计入所采用的 RPN 目标检测器和 KLT 目标跟踪器进行预处理步骤所需的时间. 所有的实验均在一个配

备单个 NVIDIA RTX 2080Ti GPU 与 Intel Core i7-7800X CPU 的服务器环境中进行. 在测试过程中, 为确保一致性, 各数据集中视频帧统一被调整为  $256 \times 256$  像素的灰度图像格式, 并且批大小被设置为 1.

#### 4.9 场景特征分析

为了验证场景特征中前景目标对本文方法产生的性能影响, 本文采用一种视频运动平均算法<sup>[28]</sup> 移除视频中移动的前景目标, 仅保留静态背景作为场景建模的输入, 这一过程前后的对比示例帧分别展示在图 9(a)和图 9(b)中. 在半监督的 Avenue 数据集上, 本文用去除前景目标的图像 (w/o moving object) 和未去除前景目标的图像 (w/moving object) 作为场景的输入, 得到的结果见表 6. 从表 6 的数据分析可得, 在以去除前景目标和未去除前景目标两种不同类型的整帧图像作为场景特征输入时, 本文方法获得近似的检测精度. 实验结果证明了视频帧中的前景目标对于场景特征提取的影响相对有限.

表 5 本文方法在 Subway、Avenue 和 ShanghaiTech 数据集上的检测时间

数据集	测试集视频帧数/帧	测试时间/s	测试速度/(帧/s)
Subway (Entrance)	116 524	4937	23.6
Subway (Exit)	64 901	2785	23.3
Avenue	15 324	669	22.9
ShanghaiTech	42 883	1880	22.8



(a) 去除前景前的视频帧



(b) 去除前景后的视频帧

图 9 去除前景前(a)和去除前景后(b)的视频帧

表 6 场景特征去除前景目标前后本文方法的 AUC 结果

训练方式	方法	AUC/%
半监督	Ours w/o moving object	91.2
	Ours w/moving object	91.4

#### 4.10 伪标签生成方法分析

为了评估本文所提方法对不同伪标签生成策略的稳健性,本文进行了一项替代实验,在 Avenue 数据集上用 K-means 聚类算法替代原有的 EIF 方法生成伪标注.具体来说,本文同样以帧级表示  $[\sum_k f_k^t / K, s^t]$  为输入,然后使用 K-means 算法进行聚类(聚类中心根据经验设为 10),接下来本文按照每个样本到最近聚类中心的距离进行排序,通过选择前  $M_1$  个样本和后  $M_2$  个样本构建伪标注正常数据集和异常数据集(与 EIF 方法生成伪标签过程类似).

从表 7 中的数据分析可得,尽管直接使用 EIF 和 K-means 方法进行异常检测的精度表现差异显著,然而以它们作为伪标签生成方法嵌入到本文提出的框架中时,本文方法却取得了近似的异常检测精度.这一现象证实了本文方法所采用的自训练策略对于伪标签生成方法的敏感度较低,能够在一定程度上抵消由伪标注准确性不足导致的负面影响.

表 7 使用不同伪标签生成方法的实验结果比较

训练方式	方法	AUC/%
无监督	EIF	78.9
	K-means	66.4
	Ours w/EIF	90.0
	Ours w/K-means	89.7

#### 4.11 自训练策略分析

本文使用自训练策略进行多轮迭代的数据伪标签生成和训练模型微调.本文在 Avenue 数据集上分别进行迭代轮数为 1、2、3、4 的实验,结果见表 8.可以看出:(1)随着迭代轮数的增加,本文方法的性能在逐渐提升;(2)第 2 轮迭代比第 1 轮迭代性能获得了极大的提升,而随着迭代轮数增加,本文方法的性能提升效果逐渐减弱,第 4 轮迭代比第 3 轮迭代仅获得很小的提升.基于上述观察,综合考虑训练效率和模型性能,本文选择迭代轮数为 2.

表 8 执行不同迭代次数的自训练策略的实验结果比较

训练方式	迭代次数	AUC/%
无监督	1	82.2
	2	90.0
	3	90.3
	4	90.4

#### 4.12 弱监督学习的阈值参数分析

本节讨论在弱监督学习下,使用不同阈值构造三元组对异常检测所产生的影响,在 UCF-Crime 数据集上的结果见表 9.实验表明在所有备选的阈值中,选择 0.5 的阈值能得到最好的效果.分析原因如下:当阈值过低(0.3)时,标记为异常帧的比例过高,超过了 50%,这导致了较高的误报率;相反,若选择较高的异常阈值(0.7),标记为异常帧的比例降至 10%左右,从而增加了漏检率.鉴于现实场景中异常事件普遍存在短暂性和稀疏性的特点,本文选择了合适的阈值为 0.5,以确保了在三元组内标记异常帧数目既不过于密集,也不过于稀疏.

表 9 UCF-Crime 数据集上选择不同阈值实验结果比较

训练方式	阈值	AUC/%	标记异常帧的百分比/%
弱监督	0.3	75.7	55.7
	0.5	83.8	25.2
	0.7	80.2	11.4

## 5 结 论

本文提出了一种上下文建模与推理的视频异常事件检测方法.本文构建了一个端到端的上下文图推理模块和异常预测模块,从而可以充分利用包括目标、关系以及场景在内的语义上下文信息来区分异常事件.本文方法将事件的视觉上下文建模为多个迭代更新的图表示,然后通过上下文图推理,可以捕获到有助于视频异常事件检测的语义上下文信息.在 UCF-Crime、Subway、Avenue 和 ShanghaiTech 数据集上的实验表明,与现有方法相比,本文方法在无监督、半监督、弱监督和监督的学习方式中,均取得了有竞争力的性能,证明了本文方法的有效性.

本文工作主要关注了上下文图的建模和推理,然而在多组属性不同的上下文信息整合利用方面,本文只使用了一个简单且有效的注意力池化操作,通过聚合上下文信息进行帧级异常事件检测.而如何更好地整合利用多组属性不同的上下文信息,以实现更优的细粒度异常事件检测,是我们未来工作研究的方向之一.此外,本文使用了简单的循环神经网络以实现平均场近似分布的计算,然而考虑其固有的时间依赖的局限性,有效的时间感知范围相对有限.而近年来出现的诸如 Transformer 网络等先进的网络结构性能更加优异,如何利用这样性能更优异的网络改进上下文图推理是另一个未来重要的研究方向.

## 参 考 文 献

- [1] Ma Cong. Research on Key Technologies of Abnormal Behavior Analysis Based on Trajectories [Ph. D. dissertation]. Beijing Jiaotong University, Beijing, 2019 (in Chinese) (马聪. 基于轨迹的异常行为分析关键技术研究[博士学位论文]. 北京交通大学, 北京, 2019)
- [2] Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 2021, 54(2): 38:1-38:38
- [3] Chong Y S, Tay Y H. Abnormal event detection in videos using spatiotemporal autoencoder//Proceedings of the International Symposium on Neural Networks. Hokkaido, Japan, 2017: 189-196
- [4] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 733-742
- [5] Ionescu R T, Khan F S, Georgescu M, et al. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7842-7851
- [6] Li N, Chang F, Liu C. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Transactions on Multimedia*, 2021, 23: 203-215
- [7] Choi M J, Torralba A, Willsky A S. Context models and out-of-context objects. *Pattern Recognition Letters*, 2012, 33(7): 853-862
- [8] Jiang F, Yuan J, Tsafaris S A, et al. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 2011, 115: 323-333
- [9] Oh J, Kim H, Park R. Context-based abnormal object detection using the fully-connected conditional random fields. *Pattern Recognition Letters*, 2017, 98: 16-25
- [10] Pang G, Yan C, Shen C, et al. Self-trained deep ordinal regression for end-to-end video anomaly detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12173-12182
- [11] Sun C, Jia Y, Hu Y, et al. Scene-aware context reasoning for unsupervised abnormal event detection in videos//Proceedings of the ACM International Conference on Multimedia. Seattle, USA, 2020: 184-192
- [12] Feng X, Song D, Chen Y, et al. Convolutional transformer-based dual discriminator generative adversarial networks for video anomaly detection//Proceedings of the ACM International Conference on Multimedia. Virtual Event, 2021: 5546-5554
- [13] Cai R, Zhang H, Liu W, et al. Appearance-motion memory consistency network for video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Event, 2021: 938-946
- [14] Deng H, Li X. Anomaly detection via reverse distillation from one-class embedding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 9737-9746
- [15] Zhang J, Wang Z, Meng J, et al. Boosting positive and unlabeled learning for anomaly detection with multi-features. *IEEE Transactions on Multimedia*, 2019, 21(5): 1332-1344
- [16] Basharat A, Gritai A, Shah M. Learning object motion patterns for anomaly detection and improved object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8
- [17] Ionescu R T, Smeureanu S, Alexe B, et al. Unmasking the abnormal events in video//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2895-2903
- [18] Zhou S, Shen W, Zeng D, et al. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 2016, 47: 358-368
- [19] Sun C, Jia Y, Song H, et al. Adversarial 3D convolutional auto-encoder for abnormal event detection in videos. *IEEE Transactions on Multimedia*, 2021, 23: 3292-3305
- [20] Gong D, Liu L, Le V, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 1705-1714
- [21] Luo W, Liu W, Lian D, et al. Future frame prediction network for video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7505-7520
- [22] Amraee S, Vafaei A, Jamshidi K, et al. Abnormal event detection in crowded scenes using one-class SVM. *Signal, Image and Video Processing*, 2018, 12: 1115-1123
- [23] Cheng K W, Chen Y T, Fang W H. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2909-2917
- [24] Wu P, Wang W, Chang F, et al. DSS-Net: Dynamic self-supervised network for video anomaly detection. *IEEE Transactions on Multimedia*, 2024, 26: 2124-2136
- [25] Wang Y, Liu T, Zhou J, et al. Video anomaly detection based on spatio-temporal relationships among objects. *Neurocomputing*, 2023, 532: 141-151
- [26] Yang Z, Liu J, Wu Z, et al. Video event restoration based on keyframes for video anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 14592-14601
- [27] Ren S, He K, Girshick R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 39(6): 1137-1149

- [28] Supreeth H S G, Patil C M. Efficient multiple moving object detection and tracking using combined background subtraction and clustering. *Signal, Image and Video Processing*, 2018, 12(6): 1097-1105
- [29] Yang Y, Fu Z, Naqvi S M. Abnormal event detection for video surveillance using an enhanced two-stream fusion method. *Neurocomputing*, 2023, 553: 126561
- [30] Doshi K, Yilmaz Y. Towards interpretable video anomaly detection//*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, USA, 2023: 2655-2664
- [31] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials//*Advances in Neural Information Processing Systems*. Granada, Spain, 2011: 109-117
- [32] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks//*Proceedings of the IEEE International Conference on Computer Vision*. Santiago, Chile, 2015: 1529-1537
- [33] Xu D, Zhu Y, Choy C B, et al. Scene graph generation by iterative message passing//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA, 2017: 5410-5419
- [34] Hariri S, Kind MC, Brunner R J. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 33(4): 1479-1489
- [35] Wang J, Cherian A. GODS: Generalized one-class discriminative subspaces for anomaly detection//*Proceedings of the IEEE International Conference on Computer Vision*. Seoul, Republic of Korea, 2019: 8201-8211
- [36] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA, 2018: 6479-6488
- [37] Sapkota H, Yu Q. Bayesian nonparametric submodular video partition for robust anomaly detection//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New Orleans, USA, 2022: 3212-3221
- [38] Adam A, Rivlin E, Shimshoni I, et al. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2008, 30(3): 555-560
- [39] Lu C, Shi J, Jia J. Abnormal event detection at 150 FPS in MATLAB//*Proceedings of the IEEE International Conference on Computer Vision*. Sydney, Australia, 2013: 2720-2727
- [40] Luo W, Liu W, Gao S. A revisit of sparse coding-based anomaly detection in stacked RNN framework//*Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy, 2017: 341-349
- [41] Munkres J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 1957, 5: 32-38
- [42] Wang S, Zeng Y, Liu Q, et al. Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection//*Proceedings of the ACM International Conference on Multimedia*. Seoul, Republic of Korea, 2018: 636-644
- [43] Song H, Sun C, Wu X, et al. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. *IEEE Transactions on Multimedia*, 2019, 22(8): 2138-2148
- [44] Yu G, Wang S, Liu X, et al. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New Orleans, USA, 2022: 13987-13998
- [45] Jiang P, Ergu D, Liu F, et al. A Review of Yolo algorithm developments. *Procedia Computer Science*, 2022, 199: 1066-1073
- [46] Liu Z, Ning J, Cao Y, et al. Video swin transformer//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans, USA, 2022: 3202-3211
- [47] Kim J, Grauman K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA, 2009: 2921-2928
- [48] Dutta J K, Banerjee B. Online detection of abnormal events using incremental coding length//*Proceedings of the AAAI Conference on Artificial Intelligence*. Austin, USA, 2015: 3755-3761

## 附录 A.

A.1. 从正式式(1)的目标函数到式(3)中近似求解的推导过程证明. 正式式(1)为

$$\hat{h} = \arg \max_h P(h|c) \quad (1)$$

条件随机场的目标函数  $P(h|c) = 1/Z(c) \exp(-E(h|c))$  写为

$$P(h) = \frac{1}{Z} \tilde{P}(h) = \frac{1}{Z} \exp\left(-\sum_i \varphi_u(h_i) - \sum_{(i,j) \in o} \varphi_p(h_i, h_j)\right) \quad (A1)$$

平均场近似方法希望求解一个分布  $Q(h|c)$  来近似  $P(h|c)$ ,

这里的  $Q(h)$  满足  $Q(h) = \prod_i Q_i(h_i)$ , 因此只需要最大化式(A2)

中 KL 散度就可以实现近似求解, 式(A2)如下:

$$\begin{aligned} D_{KL}(Q(h_i) \| P(h_i)) &= \sum_i Q(h_i) \log\left(\frac{Q(h_i)}{P(h_i)}\right) \\ &= -\sum_i Q(h_i) \log P(h_i) + \sum_i Q(h_i) \log Q(h_i) \\ &= -\mathbf{E}_{U \sim Q}[\log P(U)] + \mathbf{E}_{U \sim Q}[\log Q(U)] \\ &= -\mathbf{E}_{U \sim Q}[\log \tilde{P}(U)] + \mathbf{E}_{U \sim Q}[\log Z] + \sum_i \mathbf{E}_{U_i \sim Q_i}[\log Q(U_i)] \\ &= \mathbf{E}_{U \sim Q}[E(U)] + \sum_i \mathbf{E}_{U_i \sim Q_i}[\log Q(U_i)] + \log Z \quad (A2) \end{aligned}$$

式(A2)第三到第四行是能量函数  $E$  的等价代换形式推导, 并且因为  $Z$  只和  $P$  有关, 所以可以将它从期望中提取

出来. 除此之外, 考虑到边际概率和为 1,  $Q$  的约束条件为

$$\sum_i Q_i(h_i) = 1.$$

根据拉格朗日乘法, 构造函数(A3)

$$L = \mathbf{E}_{U \sim Q} [E(U)] + \sum_i \mathbf{E}_{U_i \sim Q_i} [\log Q(U_i)] + \log Z + \lambda \left( \sum_i Q_i(h_i) \right) \quad (\text{A3})$$

求导可得

$$\frac{\partial L}{\partial Q_i(h_i)} = \varphi_u(h_i) + \sum_{(i,j) \in o} \mathbf{E}_{U_j \sim Q_j} [\varphi_p(h_i, U_j)] + \log Q_i(h_i) + \lambda \quad (\text{A4})$$

令导数为 0, 可得式(A5)和(A6)如下:

$$\log Q_i(h_i) = -\varphi_u(h_i) - \sum_{(i,j) \in o} \mathbf{E}_{U_j \sim Q_j} [\varphi_p(h_i, U_j)] - \lambda \quad (\text{A5})$$

$$Q_i(h_i) = \exp(-\varphi_u(h_i) - \sum_{(i,j) \in o} \mathbf{E}_{U_j \sim Q_j} [\varphi_p(h_i, U_j)] - \lambda) \quad (\text{A6})$$

从而得到正文中式(3)的推导公式

$$Q_i(h_i) = \frac{1}{Z_i} \exp(-\varphi_u(h_i) - \sum_{(i,j) \in o} \mathbf{E}_{U_j \sim Q_j} [\varphi_p(h_i, U_j)]) \quad (\text{A7})$$

**A.2.** 分布  $Q(h)$  的迭代求解算法. 受到已有工作<sup>[31]</sup>的启

发, 二元能量  $\varphi_p(h_i, h_j)$  可以建模为加权高斯的函数, 即

$$\varphi_p(h_i, h_j) = \mu(h_i, h_j) \sum_m \omega^m k_G^m(h_i, h_j) \quad (\text{A8})$$

其中  $k_G^m$  是应用在  $h_i, h_j$  上的高斯核函数,  $\omega^m$  是线性加和权重, 函数  $\varphi_p(h_i, h_j)$  是兼容性函数, 表示不同最优表示  $\hat{h}_i$  和  $\hat{h}_j$  之间的兼容性. 式(A7)左右两边都有函数  $Q_i(h_i)$ , 因此可以先初始化一个  $Q_i(h_i)$ , 然后不断迭代代入, 直到目标 KL 散度小于阈值. 将式(A8)代入到式(A7)中可以得到算法 A.

**算法 A.**  $Q(h)$  的迭代求解算法.

初始化  $Q_i(h_i) \leftarrow 1/Z_i \exp(-\varphi_u(h_i))$ ;

循环直到收敛:

$$\text{消息传递 } \tilde{Q}_i^m(l) \leftarrow \sum_{(i,j) \in o} k_G^m(h_i, h_j) Q_j(l);$$

$$\text{兼容性转换 } \tilde{Q}_i(h_i) \leftarrow \sum_{h_i \in b} \mu^m(h_i, l) \sum_m \omega^m \tilde{Q}_i^m(l);$$

$$\text{更新 } Q_i(h_i) \leftarrow \exp(-\varphi_u(h_i) - \tilde{Q}_i(h_i));$$

$$\text{归一化 } Q_i(h_i);$$

结束循环.

## 附录 B.

网络结构和特征提取的实现细节, 包括上下文图表示的视觉特征提取网络、上下文图推理的循环神经网络和异常预测的注意力池化网络.

(1) 上下文图表示的视觉特征提取网络. 本文将视频帧调整为  $256 \times 256$  像素的灰度图像. 遵循文献<sup>[33]</sup>方法, 本文采用在 MS-COCO 数据集上预训练的区域备选网络 (Region Proposal Network, RPN), 以每 10 帧的间隔从图像中抽取其目标边界框集合  $B'$ , 并依据 RPN 输出的置信度保留前 10 个最可信的边界框. 随后, 本文使用 KLT (Kanade-Lucas-Tomasi) 跟踪器追踪接下来连续 9 帧的对应目标边界框. 根据正文所述操作进一步得到空间关系和场景的边界框集合. 接下来, 本文将每一帧图像通过在 MS-COCO 数据集上预训练的 VGG-16 模型来生成整幅图像的特征图 (feature map). 根据目标、空间关系和场景的边界框集合, 本文用感兴趣区域池化层 (Region of Interest Pooling Layer, ROI Pooling Layer) 提取维度为 512 的视觉特征, 作为相应的目标表示  $f_k^i \in \mathbb{R}^{512}$ 、空间关系表示  $r_{k,l}^i (k \neq l) \in \mathbb{R}^{512}$  和场景表示  $s_k^i \in \mathbb{R}^{512}$ ; 随后用可学习的参数矩阵  $W_d \in \mathbb{R}^{512 \times 1024}$  得到降维后的时间关系表示  $r_{k,k}^i = W_d[f_k^i, f_k^{i+1}] \in \mathbb{R}^{512}$ ,  $[\cdot, \cdot, \cdot]$  表示向量拼接.

(2) 上下文图推理的循环神经网络. 本文在迭代更新上下

文图表示时, 使用了三个独立的循环神经网络, 包括  $RNN_{obj}$ 、 $RNN_{rel}$  和  $RNN_{sce}$ . 为了实现高效时序建模且减少参数量, 本文选用门控循环单元 (Gated Recurrent Unit, GRU) 这一典型的 RNN 网络, GRU 的输入和输出的维度均为 512. 如正文图 4 所示, 这些 RNN 在首次迭代时以上下文的视觉特征  $f_k^i, r_{k,l}^i, s_k^i$  为输入; 接下来迭代时的输入是来自前一步骤的其他 RNN 单元的聚合消息; 最终输出更新后的上下特征  $\hat{h}_k^i \in \mathbb{R}^{512}$ 、 $\hat{h}_{k,l}^i \in \mathbb{R}^{512}$  和  $\hat{h}^i \in \mathbb{R}^{512}$ .

(3) 异常预测的注意力池化网络. 本文通过注意力池化网络聚合迭代更新后的上下特征  $\hat{h}_k^i, \hat{h}_{k,l}^i$  和  $\hat{h}^i$ , 得到了帧级异常分数. 本文首先使用一个可学习参数  $W_p \in \mathbb{R}^{1 \times 512}$  的网络层对目标表示进行池化, 得到池化目标表示  $f^i \in \mathbb{R}^{512}$ , 正文的式(9)展示了这一动态池化过程. 随后, 本文采用另一个具备可学习参数  $W_q \in \mathbb{R}^{1 \times 512}$  的网络层, 以类似的方式对特征表示进行池化, 得到池化目标表示  $r^i \in \mathbb{R}^{512}$ . 本文将池化目标表示  $f^i$ 、池化关系表示  $r^i$  和更新后的场景表示拼接起来得到  $[f^i, r^i, \hat{h}^i] \in \mathbb{R}^{1536}$ , 然后将它们输入到一个两层的全连接网络中, 得到异常的概率分布  $P(y^i | \cdot)$ . 这一过程如原文的式(10)所示, 式中  $W_o \in \mathbb{R}^{1536 \times 256}$  和  $W_y \in \mathbb{R}^{256 \times 1}$  是全连接网络的可学习参数,  $\phi$  和  $\sigma$  分别是 Sigmoid 和 SoftMax 激活函数.



**SUN Che**, Ph. D. His research interests include computer vision, machine learning, and video analysis.

**WU Yu-Wei**, Ph. D., tenured associate professor. His research interests include computer vision, machine learning, and multimedia analysis.

**JIA Yun-De**, Ph. D., professor. His research interests include computer vision, artificial intelligence, and cognitive computing and systems.

## Background

This paper focuses on the research on abnormal event detection in videos, which is a hot topic in computer vision (CV) and machine learning (ML). Many recent methods have paid more attention to learning event patterns based on appearance and motion features that are usually extracted from an image or an object region in the image. They would ignore the influence of the visual context information on anomaly prediction. The visual context is one of the important bases for discriminating abnormal events, and it is reflected in various objects and their relationships as well as scene types surrounding the event in a video. Mining context information beyond image-level and object-level features is beneficial to anomaly prediction. To this end, we can discriminate abnormal events by the proposed context modeling and reasoning method in this paper.

Prior to our work, existing efforts of context-based abnormal event detection, Choi M J et al., Jiang F et al., Oh J et al., etc., manually pre-define the collections of context based on the human experience. They develop context models of specific relationships among objects in specific video scenes for discriminating anomalies, such as support relationships, co-occurrence relationships, geometric

relationships, etc. The correctness and completeness of the collections are crucial to the performance of abnormal event detection. Unfortunately, it is impossible to manually pre-define the collections that take all possible context information in videos into account, because in many cases the definition of context-related behaviors is diverse, constantly changing, and unpredictable.

In this work, we design a context modeling and reasoning method that automatically learns context information from data rather than manually pre-define contextual contents, which meliorates the insufficient in discriminating context-based abnormal events. Our method mines high-level context information from low-level visual features of data, which bridges the semantic gap between visual context and the meaning of abnormal events. The experiment results show that our method yield better detection results of context-related abnormal events in various scenes.

This work was supported in part by the General Program of Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006 and the Key Research Platform and Program of Guangdong Provincial Department of Education for Ordinary Universities under Grant No. 2023ZDZX1034.