ISSN 1003-501X CN 51-1346/O4

光电工程

Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

改进时空图卷积网络的视频异常检测方法

张红民，颜鼎鼎，田钱前

引用本文:

张红民，颜鼎鼎，田钱前. 改进时空图卷积网络的视频异常检测方法[J]. 光电工程，2024， **51**(5): 240034.

Zhang H M, Yan D D, Tian Q Q. Improved spatio-temporal graph convolutional networks for video anomaly detection[J]. *Opto-Electron Eng*, 2024, **51**(5): 240034.

https://doi.org/10.12086/oee.2024.240034

收稿日期: 2024-02-01; 修改日期: 2024-04-09; 录用日期: 2024-04-10

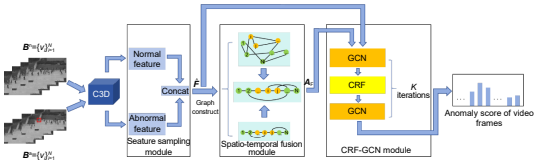
相关论文

Opto-Electronic Engineering
光 电 工 程

Article
2024 年, 第 51 卷, 第 5 期

DOI: 10.12086/oee.2024.240034

改进时空图卷积网络的视频异常检测方法



张红民*，颜鼎鼎，田钱前

重庆理工大学电气与电子工程学院，重庆 400054

摘要： 为了对异常事件中对象的时空相互作用进行精准捕捉，提出一种改进时空图卷积网络的视频异常检测方法。在图卷积网络中引入条件随机场，利用其对帧间特征关联性的影响，对跨帧时空特征之间的相互作用进行建模，以捕捉其上下文关系。在此基础上，以视频段为节点构建空间相似图和时间依赖图，通过二者自适应融合学习视频时空特征，从而提高检测准确性。在 UCSD Ped2、ShanghaiTech 和 IITB-Corridor 三个视频异常事件数据集上进行了实验，帧级别 AUC 值分别达到 97.7%、90.4% 和 86.0%，准确率分别达到 96.5%、88.6% 和 88.0%。

关键词： 视频异常检测；图卷积网络；条件随机场

中图分类号： TP391　　**文献标志码：** A

张红民，颜鼎鼎，田钱前. 改进时空图卷积网络的视频异常检测方法 [J]. 光电工程，2024，**51**(5): 240034

Zhang H M, Yan D D, Tian Q Q. Improved spatio-temporal graph convolutional networks for video anomaly detection[J]. *Opto-Electron Eng*, 2024, **51**(5): 240034

Improved spatio-temporal graph convolutional networks for video anomaly detection

Zhang Hongmin*, Yan Dingding, Tian Qianqian

School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract: An improved spatio-temporal graph convolutional network for video anomaly detection is proposed to accurately capture the spatio-temporal interactions of objects in anomalous events. The graph convolutional network integrates conditional random fields, effectively modeling the interactions between spatio-temporal features across frames and capturing their contextual relationship by exploiting inter-frame feature correlations. Based on this, a spatial similarity graph and a temporal dependency graph are constructed with video segments as nodes, facilitating the adaptive fusion of the two to learn video spatio-temporal features, thus improving the detection accuracy. Experiments were conducted on three video anomaly event datasets, UCSD Ped2, ShanghaiTech, and IITB-Corridor, yielding frame-level AUC values of 97.7%, 90.4%, and 86.0%, respectively, and achieving accuracy rates of 96.5%, 88.6%, and 88.0%, respectively.

Keywords: video anomaly detection; graph convolutional network; conditional random field

收稿日期：2024-02-01；修回日期：2024-04-09；录用日期：2024-04-10

基金项目：国家自然科学基金资助项目 (61901068)；重庆市自然科学基金面上项目 (cstc2021 jcyj-msxmX0525，CSTB2022NSCQ-MSX0786，CSTB2023NSCQ-MSX0911)；重庆市教委科学技术研究项目 (KJQN202201109)

*通信作者：张红民，hmzhang@cqut.edu.cn。

1 引 言

视频监控系统越来越广泛地被用于公共场所，并在维护社会安全和稳定方面发挥了重要作用[1]。然而，异常视频的收集与标记受制于主观因素，导致视频数据仅含有视频级别的标签而缺乏详细信息，限制了对视频的智能化分析，特别是在异常检测领域，需要更丰富的数据信息以提高模型性能。当前研究将基于这种弱监督数据进行的视频异常检测问题视为多示例学习[2]，但这种方法在处理联系性较强的异常事件时效果不佳。鉴于上述情况，Zhou 等[3] 提出多示例学习中各个示例并不是独立同分布的，他们之间存在一定的联系，学习并利用这种联系将更好地表达视频的动态性特征，更有利于多样性视频的异常事件检测。视频数据是一种典型的时空数据，视频中异常事件所展现的时空特征具有显著的相关性，可通过时序信息进行显式建模[4]，引入时间角度和空间角度的图结构去构建视频中各个片段之间的联系，但传统的卷积操作无法直接运用到图中[5]。图卷积神经网络 (graph convolutional network，GCN) 虽然具有较好的结构信息获取能力[6]，但在捕捉相邻帧中对象之间的内在关系方面仍存在不足，尤其在应对视频序列中帧间复杂时空依赖关系时表现欠佳。为了在图结构下更合理地对视频片段的时空关联性进行建模，进而有效检测与定位视频异常，本文提出一种改进时空图卷积网络的视频异常检测方法。将视频中的各个片段视作节点，并构建了两个关键的图模型，空间相似图和时间依赖图。在考虑片段间时空联系的基础上，通过自适应融合的方式学习视频特征。由于异常事件可以通过多个对象之间的时空相互作用形成，利用条件随机场 (conditional random field，CRF) 良好的图形建模优势，在 GCN 模型中引入 CRF 层，对跨帧的时空特征之间的相互作用进行建模，以捕捉其上下文关系，从而提高模型的检测准确性。

2 相关工作

视频异常检测是一个具有挑战性的研究问题。传统方法手工提取特征进行训练与推断，依赖特征选择且对场景具有局限性，为适应多样化的场景，近年来出现了一些针对复杂场景的弱监督视频异常检测技术。Sultani 等[7] 首次将多示例学习 (multiple instance learning，MIL) 引入视频异常检测中，将视频按照是否异常分为正常包和异常包，并将视频片段视为实例，通过深度学习异常排序模型，预测视频片段的异常分数来判定是否异常。Zhang 等[8] 基于正负包内各个示例的差异提出包内损失，并利用时序卷积神经网络进行时序上的关联，但该方法仅扩大了正常与异常的差异分化，对于趋向中性的视频片段效果不佳。Li 等[9] 为了提高训练初期异常包内异常片段选择的准确率，基于多序列学习及其排序损失，设计了一个基于 Transformer 的多序列学习网络，并在推理阶段使用视频级异常分类概率抑制片段级异常得分的波动。之前的 MIL 方法往往忽略了相对于正常特征异常特征应该具有持续的显著性，对此 Liang 等[10] 提出一种时空特征融合增强学习方法来解决该问题。

时空图神经网络作为一种融合时间序列特征的属性图网络，具有独特的优势。它能够有效地捕捉图结构中时间域和空间域的特征信息，使得网络能够更全面地理解复杂的时空动态变化[11]。Zhou 等[12] 针对视频异常事件发生的时空特征相关性进行有效融合，形成时空融合图卷积。Purwanto 等[13] 在时空图卷积的基础上，结合关系感知特征提取和自注意力的条件随机场，处理特征的局部和非局部关系，采用对比式多实例学习方案增强正常和异常实例之间的差异；Mu 等[14] 采用空间相似性和时间一致性，基于自适应加权方法构建基于时空图的 CNNs(STGCNs)。Liu 等[15] 针对和正常行为相似的异常行为提出级联多层次上下文内容分析模块 (CMC)，采用时间感知图卷积网络和非本地神经网络来聚合本地和非本地视频剪辑的上下文特征；Cheng 等[16] 设计了一种无监督时空图卷积网络 (STGCN) 增强的流帧预测模型，通过学习运动正态性和外观正态性的潜在表示来增强模型学习时空正态性。Li 等[17] 提出了一种时空图引导的全局注意网络 (STG3A)，能够捕获一个框架内的空间上下文和整个所有框架内的时间上下文，并学习时空上下文信息。

然而，现有方法往往局限于单一视角，要么专注于时空关联性，要么只考虑到局部和非局部特征的关系。在充分挖掘视频片段中复杂时空关系的同时，这种单一视角的研究框架未能全面捕捉到局部和非局部特征之间的有效关联。因此，本文采用了一种综合考虑跨帧时空特征关联性及其上下文关系的策略，通过引入条件随机场对时空图卷积网络进行改进，使其更有效地表达视频内在的复杂关联。

3 改进时空图卷积网络模型

本文将视频异常事件检测视为 MIL 问题。取正常视频$\boldsymbol{B}^{\mathrm{n}}=\{v_i\}_{i=1}^{N}$作为负包，异常视频$\boldsymbol{B}^{\mathrm{a}}=\{v_i\}_{i=1}^{N}$作为正包，包内的每一个视频片段代表一个示例，每个包内都含有一个视频级别的标签，其中 1 代表异常，0 代表正常。本文所提方法的整体网络结构如图 1 所示，共分为 3 个部分：1) 特征采样模块。通过 C3D 网络对正常视频和异常视频进行特征提取，并进行采样和拼接，作为图构建时的输入。2) 时空融合模块。分别从时间角度和空间角度进行图构建，通过自适应的方式融合得到时空图，作为图卷积计算的算子。3) CRF-GCN 模块。将采样得到的视频特征与时空融合图的邻接矩阵送入 CRF-GCN 模块，经过 K 次迭代计算，最终得到视频帧异常得分。

3.1 特征采样模块

首先以 C3D 为特征提取器分别提取正常视频$\boldsymbol{B}^{\mathrm{n}}$和异常视频$\boldsymbol{B}^{\mathrm{a}}$的特征表示$\boldsymbol{F}\in\mathbb{R}^{N\times d}$，其中 d 表示每个视频片段的特征维度。为避免对视频进行细粒度切片导致的参数过拟合问题[18]，在训练时采用稀疏连续采样策略[19]，从$\boldsymbol{F}\in\mathbb{R}^{N\times d}$中采样 m 个特征子包，每个子包由 t 个连续的视频特征$\{\boldsymbol{F}_i\}_{i=1}^{t}$组成，经采样后得到每个视频的特征$\hat{\boldsymbol{F}}\in\mathbb{R}^{M\times d}$，其中$M=m\times t$。其中，超参数 t 代表着视频内异常的持续时间，而超参数 m 与视频的总长度相关。

3.2 时空融合模块

视频片段隐含着时间连续性和空间相似性。通过构建空间相似图和时间依赖图，在考虑视频片段间时空联系的基础上，基于自适应融合的方式学习视频特征。

3.2.1 空间相似图

视频片段之间存在空间上的相似性联系，一般来说，具有相似特征的视频片段应该分配近似的异常分数。因此，在视频片段特征上构建空间相似图$G_{\mathrm{S}}=(V,E_{\mathrm{S}})$，学习潜在特征相似性。节点集$V=\left\{v_i=f_i,f_i\in\hat{\boldsymbol{F}}\right\}$，其中$\{f_i\}_{i=1}^{M}$为 $\boldsymbol{F}$ 的第 i 行，即第 i 个视频片段。E_{S}表示视频片段节点之间的特征相似性。邻接矩阵$\boldsymbol{A}_{\mathrm{S}}\in\mathbb{R}^{M\times M}$的计算如式 (1) 所示：

$$\boldsymbol{A}_{\mathrm{S}_{(i,j)}}=\exp\left(f_i\cdot f_j-|f_i|^2\right), \tag{1}$$

其中：内积用于计算第 i 个视频片段和第 j 个视频片段之间的特征相似性，指数函数$\exp(\cdot)$用于将$\boldsymbol{A}_{S_{(i,j)}}$限制在范围$(0,1]$内。考虑到无向图在图传播节点信息的优势，对邻接矩阵进行对称化$\boldsymbol{A}_{\mathrm{S}}=(\boldsymbol{A}_{\mathrm{S}}+(\boldsymbol{A}_{\mathrm{S}})^{\mathrm{T}})/2$，采用转置矩阵相加使图转变为无向图$G_{\mathrm{S}}$，加强相似节点间的联系。

3.2.2 时间依赖图

在时间维度上相邻的视频片段应该有相近的异常分数，因为视频上下文通常不会在短时间内发生剧烈变化。因此，构造时间依赖图$G_{\mathrm{T}}=(V,E_{\mathrm{T}})$来获取相邻时间片段的时间连续性。节点集$V=\{v_i,i=1,...,M\}$代表以视频片段为节点的集合，$E_{\mathrm{T}}$表示第 i 个视频片段和第 j 个视频片段在时间上的相似性。与空间相似图不同，时间依赖图只与视频片段的时间顺序相关。邻接矩阵$\boldsymbol{A}_{\mathrm{T}}\in\mathbb{R}^{M\times M}$的计算如式 (2) 所示：

$$\boldsymbol{A}_{\mathrm{T}_{(i,j)}}=\begin{cases}1, & i=j\\ \dfrac{1}{|i-j|}, & i\neq j\end{cases}. \tag{2}$$

当$i=j$时，代表两个视频片段相同，此时时间关联性为 1。当$i\neq j$时，用第 i 个视频片段与第 j 个视频片段之间的时间距离反比来衡量其时间关联性。

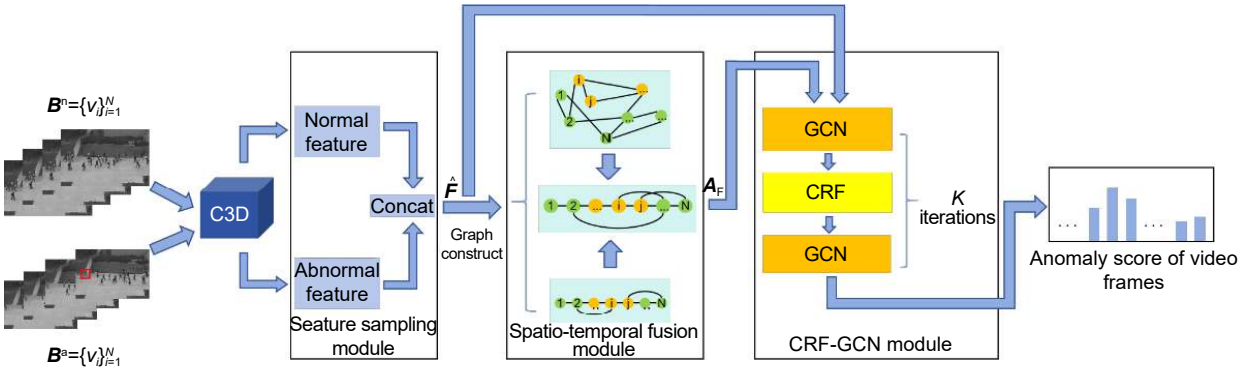


图 1 改进时空图卷积网络模型框架

Fig. 1 Improved spatio-temporal graph convolutional network model framework

3.2.3 自适应时空融合图

空间相似图描述了视频片段内的异常事件在特征上的内在关联关系，时间依赖图描述了视频片段在时间序列上的远近程度，有助于确定异常事件与正常事件之间的边界。因此，对于准确定位异常事件而言，单一的空间相似图或单一的时间依赖图存在一定的局限性。参照 Zhou 等[12] 所提方法，本文提出一种改进的自适应方式融合空间相似图与时间依赖图。方式如下：给定邻接矩阵$\boldsymbol{A}_{\mathrm{S}}, \boldsymbol{A}_{\mathrm{T}} \in \mathbb{R}^{M\times M}$，分配待学习的权重$\boldsymbol{W}_{\mathrm{S}} \in \mathbb{R}^{M\times M}$与$\boldsymbol{W}_{\mathrm{T}} \in \mathbb{R}^{M\times M}$和偏置$\boldsymbol{b}_{\mathrm{S}} \in \mathbb{R}^{M\times M}$与$\boldsymbol{b}_{\mathrm{T}} \in \mathbb{R}^{M\times M}$，得到两个邻接矩阵的概率矩阵，如式 (3) 所示：

$$\begin{cases} \boldsymbol{p}_{\mathrm{S}} = \sigma(\mathrm{softmax}(\boldsymbol{W}_{\mathrm{S}}\boldsymbol{A}_{\mathrm{S}} + \boldsymbol{b}_{\mathrm{S}})) \\ \boldsymbol{p}_{\mathrm{T}} = \sigma(\mathrm{softmax}(\boldsymbol{W}_{\mathrm{T}}\boldsymbol{A}_{\mathrm{T}} + \boldsymbol{b}_{\mathrm{T}})) \end{cases}, \tag{3}$$

其中：σ为 sigmoid 激活函数，将概率映射到$[0,1]$的分布范围。通过注意力机制对邻接矩阵加权，最终得到时空融合图$\boldsymbol{A}_{\mathrm{F}}$，如式 (4) 所示：

$$\boldsymbol{A}_{\mathrm{F}} = \left(\boldsymbol{p}_{\mathrm{S}} + (\boldsymbol{p}_{\mathrm{S}})^{\mathrm{T}}\right) * \boldsymbol{A}_{\mathrm{S}} + \left(\boldsymbol{p}_{\mathrm{T}} + (\boldsymbol{p}_{\mathrm{T}})^{\mathrm{T}}\right) * \boldsymbol{A}_{\mathrm{T}}\,. \tag{4}$$

融合图$G_{\mathrm{F}} = (V, E_{\mathrm{F}})$为概率矩阵与邻接矩阵的点积，其中节点集 V 代表以视频片段为节点的集合，边集E_{F}的邻接矩阵为$\boldsymbol{A}_{\mathrm{F}} \in \mathbb{R}^{M\times M}$，归一化后得到矩阵$\hat{\boldsymbol{A}}_{\mathrm{F}}$，如式 (5) 所示：

$$\hat{\boldsymbol{A}}_{\mathrm{F}} = \tilde{\boldsymbol{D}}_{\mathrm{F}}^{\frac{1}{2}} \tilde{\boldsymbol{A}}_{F} \tilde{\boldsymbol{D}}_{\mathrm{F}}^{\frac{1}{2}}\,, \tag{5}$$

其中：$\tilde{\boldsymbol{A}}_{\mathrm{F}} = \boldsymbol{A}_{\mathrm{F}} + \boldsymbol{I}$为矩阵自增，$\tilde{\boldsymbol{D}}_{\mathrm{F}}$为时空融合图自增后的度矩阵。

3.3 CRF-GCN 模块

给定图$G = \{\boldsymbol{A}, X\}$，GCN 的单层传播公式为$\boldsymbol{H}^{(l+1)} = \sigma\left(\hat{\boldsymbol{A}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l+1)}\right)$。其中，$\sigma$为 sigmoid 激活函数，$\boldsymbol{H}^{(l)}$为 GCN 在第 l 层的输出，$\boldsymbol{H}^{(0)} = \hat{\boldsymbol{F}}$，$\boldsymbol{W}^{(l+1)} \in \mathbb{R}^{M\times M}$为第$l+1$层的可学习权重矩阵。$\hat{\boldsymbol{A}}\boldsymbol{H}^{(l)}$代表聚合邻接节点的特征，$\boldsymbol{H}\boldsymbol{W}^{(l+1)}$表示通过非线性变换进行特征提取。节点间的边承载着不同节点之间的相似性关系。

由于异常事件可以通过多个对象之间的时空相互作用形成，因此了解相邻帧中对象之间的内在关系对于异常检测具有重要意义。CRF[20] 作为基于无向图的概率判别模型，具有良好的图形建模优势，可通过引入节点间的关联性来影响节点的特征表示，并允许在整个图像序列上进行联合概率推断。因此引入 CRF 层以扩展 GCN 模型，使得 GCN 模型更加注重节点之间的关联性和时空信息的整合，捕捉视频序列中帧间的复杂时空依赖关系。改进后的 CRF-GCN 模块如图 2 所示。

CRF 的定义如下：给定输入数据x_i，CRF 通过最大化条件概率$P(y_i|x_i) = z^{-1}(x_i)\exp(-E(y_i|x_i))$来为其分配标签$y_i$。其中$E(y_i|x_i)$表示能量函数，$z(x_i)$为归一化分配函数[20]。对于一个全连通的 CRF 模型，总能量如式 (6) 所示[21]：

$$E(y_i|x_i) = \sum \varphi_u(y_i|x_i) + \sum_j \varphi_p(y_i, y_j|x_i, x_j)\,, \tag{6}$$

其中：$\varphi_u(y_i|x_i)$为一元能量，表示在给定输入为x_i时分配标签y_i所需的损失，仅考虑了单个输入的信息；$\varphi_p(y_i, y_j|x_i, x_j)$为二元能量，表示在考虑输入$x_i$与$x_j$相关性的前提下分配标签$y_i$所需的损失，考虑了与$x_i$有关的上下文信息。

考虑一组节点标签$X = \{0,1\}$，其中 0 代表正常，1 代表异常。将 GCN 在第 l 层的输出$\boldsymbol{H}^{(l)}$视为随机变量$\left\{\boldsymbol{H}_i^{(l)}\right\}$，其中$\boldsymbol{H}_i^{(l)}$为$\boldsymbol{H}^{(l)}$的第 i 行，对应图 G 中第 i 个节点的标签预测。这些随机变量以学习到的视频特征$\{\boldsymbol{F}_i\}$为条件。基于上述表述，得到 CRF-GCN 模型如下，其中$E(\boldsymbol{H}|\boldsymbol{F})$为总能量函数，由一元能量函数$\varphi_u$和二元能量函数$\varphi_p$两部分组成：

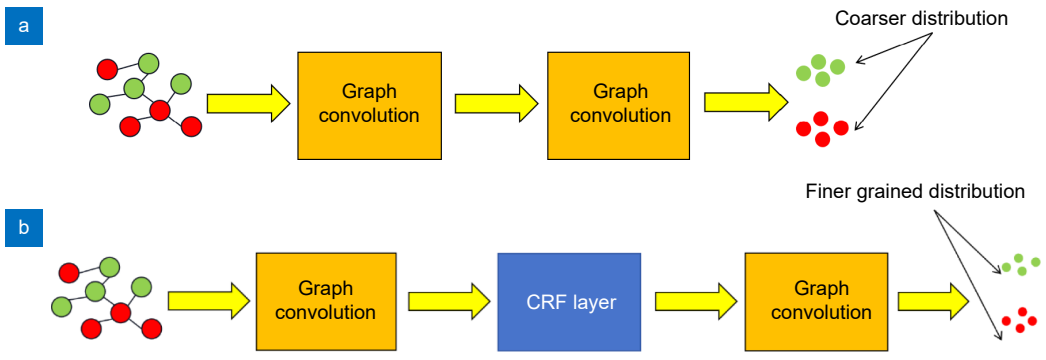


图 2 GCN 模块与 CRF-GCN 模块对比。(a) GCN 模块；(b) CRF-GCN 模块

Fig. 2 Comparison between GCN module and CRF-GCN module. (a) GCN module; (b) CRF-GCN module

$$P(\boldsymbol{H}|\boldsymbol{F})=\frac{1}{Z(\boldsymbol{F})}\exp(-E(\boldsymbol{H}|\boldsymbol{F}))\ . \tag{7}$$

对第 i 个节点的标签预测的一元能量可以通过单层 GCN 传播公式来确定，$\varphi_u(\boldsymbol{H}_i^{l+1}|\boldsymbol{F}_i)=\sigma\left(\hat{\boldsymbol{A}}\boldsymbol{H}_i^{(l)}\boldsymbol{W}^{(l+1)}\right)$，它代表将标签分配给第 i 个节点的代价。二元能量函数$\varphi_p(\boldsymbol{H}_i,\boldsymbol{H}_j|\boldsymbol{F}_i,\boldsymbol{F}_j)=\sum_{i\neq j}u(\boldsymbol{H}_i,\boldsymbol{H}_j)\hat{p}(f_i,f_j)m_p(f_i)$，$m_p(f_i)$是一个线性前馈层，将节点表示转换为标签预测。$u(\boldsymbol{H}_i,\boldsymbol{H}_j)$是兼容性函数，可以由 Potts 模型决定，当$i\neq j$时，$u(\boldsymbol{H}_i,\boldsymbol{H}_j)=1$，否则为 0。$\hat{p}(f_i,f_j)=\sum_{\tau\in K}\sum_{j\in C_j^\tau}\omega_\tau g_{ij}s$，其中，$g_{ij}$表示节点 i 与节点 j 之间的相似性。因此，相似的视频片段将被鼓励具有相同的标签。对一元能量函数和二元能量函数进行加权，得到最终总的能量函数，如式 (8) 所示：

$$\begin{aligned}E\left(\boldsymbol{H}_i^{(l)}|\boldsymbol{B}_i^{(l)}\right)=&\alpha\left(\sigma\left(\hat{\boldsymbol{A}}\boldsymbol{H}_i^{(l)}\boldsymbol{W}^{(l+1)}\right)\right)\\&+\beta\left(\sum_{i\neq j}u(\boldsymbol{H}_i,\boldsymbol{H}_j)\hat{p}(f_i,f_j)m_p(f_i)\right).\end{aligned} \tag{8}$$

CRF 通常包括大量的随机变量和潜在的依赖关系，在给定观测数据的情况下，计算所有可能的随机变量的后验概率分布通常需要处理高维空间和复杂的概率分布。为了降低计算复杂度，采用平均场近似法推导出一个易于处理和有效的更新规则。其基本思想是找到一个简单的分布$Q(\boldsymbol{H})$来近似$P(\boldsymbol{H}|\boldsymbol{F})$。这个简单的分布可以用独立边际分布的乘积表示为$Q(\boldsymbol{H})=\prod_{i=1}^{n}Q(\boldsymbol{H}_i)$。$Q(\boldsymbol{H}_i)$由一元能量和二元能量给出[22]：

$$Q(\boldsymbol{H}_i)=\frac{1}{Z_i}\exp(-(\varphi_u+\varphi_p))\ , \tag{9}$$

其中：Z_i为归一化常数[23]。与常用的将 CNN 内核堆叠以修正边际分布的推理方案不同[13]，本文提出了一种新的平均场推理算法，该算法通过将平均场推理与图卷积网络相结合，以学习节点之间的非局部关系。整体的迭代平均场推理算法流程如图 3 所示。收敛后得到的$\hat{E}$作为最终的边际分布。

3.4 损失函数

损失函数由两部分组成，包括排序损失和稀疏性约束。模型输出代表了与视频片段对应的异常分数。采用排序损失来约束模型学习异常视频中的最高得分与正常视频中的最高得分之间的差异，强化模型对异常事件的感知和区分能力：

$$L_{\text{rank}}=\max\left(0,1-\max_{i\in\boldsymbol{B}_{\text{a}}}P_i+\max_{i\in\boldsymbol{B}_{\text{n}}}Q_i\right), \tag{10}$$

其中：P_i表示异常视频$\boldsymbol{B}^{\text{a}}$中第 i 个异常视频段的异常概率值，Q_i表示正常视频$\boldsymbol{B}^{\text{n}}$中第 i 个正常视频段的异常概率值。

由于异常事件属于偶然发生事件，因此在异常视频$\boldsymbol{B}^{\text{a}}$中应当只有极少数的异常概率值分数较高。为使视频中的异常得分片段满足稀疏特征，采用稀疏性约束：

$$L_{\text{sparse}}=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{B}_i^{\text{a}}\ . \tag{11}$$

最后，得到总损失函数$L=L_{\text{rank}}+\lambda L_{\text{sparse}}$，其中$\lambda$为超参数。

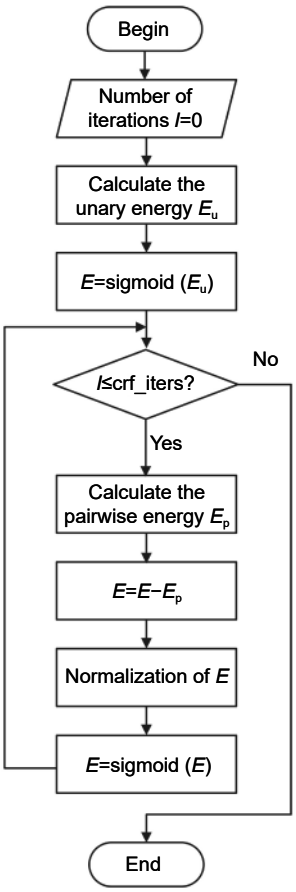


图 3 CRF-GCN 的平均场推理流程图

Fig. 3 Flowchart of mean-field inference for CRF-GCN

4 实验结果与数据分析

在三个视频异常检测数据集 UCSD Ped2[24]、ShanghaiTech[25] 和 IITB-Corridor[26] 上对所提方法进行了评估，分析了该方法的性能，并与主流方法进行了比较。数据集的详细信息如表 1 所示。此外，最后还进行了消融实验，以显示各个部分对提出方法的稳定性和重要性。

表 1 UCSD Ped2、ShanghaiTech 和 IITB-Corridor 数据集

Table 1 UCSD Ped2, ShanghaiTech and IITB-Corridor datasets

数据集	帧数	年份	标注	分辨率	异常类型
UCSD Ped2	4560	2010	Frame-level	360×240	骑自行车、小型车辆
ShanghaiTech	317398	2016	Frame-level	480×856	骑自行车、逃票、打架
IITB-Corridor	483566	2020	Frame-level	1920×1080	抗议、打斗、追逐等

4.1 数据集与评价指标

1) UCSD Ped2 数据集。Zhong 等[27] 从 Ped2 中随机提取 6 个异常视频和 4 个正常视频来构建训练集，并将所有剩余的视频分配给测试集。为了保证公平的比较，本文采用相同的数据集分区比例，重复该过程 10 次，得到平均结果。

2) ShanghaiTech 数据集，中型 VAD 数据集。总共包含 437 个视频，其中 307 个正常视频，130 个异常视频。Zhong 等[27] 重组了数据集，以使其适应弱监督设置，训练集和测试集各含 238 和 199 个视频，均包括正常和异常样本。本文应用了与文献 [27] 中相同的数据集划分。

3) IITB-Corridor 数据集，大型 VAD 数据集。对 IITB-Corridor 数据集进行新的划分，在原训练集中随机选取 166 个，原测试集中随机选取 126 个组成新的训练集，剩余的组成新的测试集，以满足弱监督设置。

4) 评价指标 AUC 值。参照已有的研究方案，采样帧级标签作为评价标准，得到接受者操作特性曲线 (receiver operating characteristic curve，ROC)，计算曲线下面积 (area under curve，AUC) 值，以 AUC 值作为最终实验结果的对比参照，AUC 值越高模型的判别效果越好。

4.2 实验参数设置

实验环境：实验服务器配置为 12 vCPU Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz，GPU 采用 RTX 3080 Ti(12 GB)，内存 90 GB。服务器采用 Ubuntu 18.04 系统，编程环境为 Python 3.8.10，CUDA 11.1，Pytorch 1.9.0。

参数设置：对视频进行重新调整，每帧大小重设为 224×224。通过在 Sports1M 数据集上预训练的 C3D 网络模型，对每个视频片段进行连续 16 帧的 RGB 图像特征提取。在此基础上，CRF-GCN 模块的迭代次数设定为 K=4。设定 dropout 率为 0.3。采用了 adagrad 优化算法，并进行了学习率衰减，初始学习率为 0.001，衰减率为 4e-5。为了更好地优化模型，UCSD Ped2、ShanghaiTech 和 IITB-Corridor 的批次大小依次设置为 4、16 和 15。在整个训练过程中共进行了 600 个 epoch 的训练。

4.3 实验结果与分析

4.3.1 方法对比分析

1) UCSD Ped2 数据集。根据相关研究在 UCSD 数据集上测试了本文所提方法，结果如表 2 所示。本文所提方法在基于图卷积网络的基础上同时考虑了特征的时间依赖性和空间关联性，取得了 97.7% 的帧级别 AUC 和 96.3% 的准确率。可以看出，文献 [28] 基于无监督的方法在进行异常检测时帧级别 AUC 与准确率均较低。从基于图的角度来看，本文所提方法的帧级别 AUC 相较于典型的示例方法[27] 高出 4.5%，相较于文献 [33] 的方法高出 0.3%，而准确率分别高出 6.2% 和 0.2%。文献 [7] 通过利用弱标记的训练视频，通过 MIL 框架学习异常，文献 [32] 在此基础上考虑了特征的时间依赖性，可以达到 96.5% 的帧级别 AUC。文献 [34] 在 Ped2 数据集上实现了与本文所提方法一致的帧级别 AUC，但由于其模型在单帧输入的基础上通过更少的迭代次数来实现较高的性能，导致其存在较高的错误检测率，因此在准确率方面低于本文所提方法。

表 2 UCSD Ped2 数据集上不同方法的对比结果

Table 2 Comparison results of different methods on UCSD Ped2 dataset

监督方式	对比方法	特征提取方式	AUC/%	准确率/%
无监督方式	Hasan的方法[28]	-	90.0	89.5
	Gong的方法[29]	-	94.1	-
	Yu的方法[30]	-	97.3	95.6
	Taghinezhad的方法[31]	Encoder	97.6	-
弱监督方式	GCN-Anomaly[27]	TSN	93.2	90.3
	Sultani的方法[7]	I3D	92.3	-
	RTFM[32]	TSN	96.5	-
	Chen的方法[33]	C3D	97.4	96.1
	Wang的方法[34]	Encoder	97.7	93.4
	本文方法	**C3D**	**97.7**	**96.5**

2) ShanghaiTech 数据集。在 ShanghaiTech 数据集上采用不同方法的结果如表 3 所示。本文所提方法在 C3D-RGB 特征上实现了 90.4% 的帧级别 AUC 与 88.6% 的准确率。与现有的基于 MIL 的模型[7, 27]进行比较，在帧级别 AUC 与准确率方面分别得到了 4.1% 与 6% 的提升。从基于图的角度来看，本文所提方法在帧级别 AUC 方面比典型的示例[12]高出 0.6%。

表 3 ShanghaiTech 数据集上不同方法的对比结果

Table 3 Comparison results of different methods on ShanghaiTech dataset

监督方式	对比方法	特征提取方式	AUC/%	准确率/%
无监督方式	Hasan的方法[28]	-	60.8	60.1
	Gong的方法[29]	-	71.2	-
	Yu的方法[30]	-	74.4	72.6
	Tur的方法[35]	3D-ResNet18	76.1	-
弱监督方式	GCN-Anomaly[27]	TSN	84.4	82.6
	Sultani的方法[7]	I3D	86.3	-
	Zhou的方法[12]	I3D	89.8	-
	Acsintoae的方法[36]	-	83.7	86.1
	Wang的方法[34]	Encoder	71.3	82.6
	本文方法	**C3D**	**90.4**	**88.6**

3) IITB-Corridor 数据集。在 IITB-Corridor 数据集上采用不同方法的结果如表 4 所示。本文所提方法在 C3D-RGB 特征上实现了 86.0% 的帧级别 AUC。

4.3.2 异常事件检测可视化结果

为了直观地显式本文所提方法的检测性能，图 4~6 分别给出了在 UCSD Ped2、ShanghaiTech 和 IITB-Corridor 三个数据集上的可视化的结果。实验在来自 UCSD Ped2 (Test003、Test012)、ShanghaiTech (04_0004、12_0173) 和 IITB-Corridor (Test000228、Train000139) 的 6 个测试视频上进行。具体来说，本文所提方法对异常片段产生的异常分数更高，对正常片段产生的异常分数更低，显示了准确的异常检测。

表 4 IITB-Corridor 数据集上不同方法的对比结果

Table 4 Comparison results of different methods on IITB-Corridor dataset

监督方式	对比方法	特征提取方式	AUC/%
无监督方式	Zeng的方法[37]	-	73.9
弱监督方式	Li的方法[38]	C3D	72.2
	Cao的方法[39]	CVAE	73.6
	Royston的方法[26]	I3D	67.1
	Majhi的方法[40]	**I3D**	84.1
	本文方法	**C3D**	**86.0**

4.3.3 算法复杂度对比分析

网络的复杂性和推理速度是评价深度学习模型的关键因素。为了进一步研究本文所提方法的性能，将其与基于图卷积的方法和基于其他框架的方法进行比较，分别使用了乘法累积运算 (MACs) 和模型参数量 (Params) 来量化时间和空间的复杂性，结果如表 5 所示。每个片段的尺寸为 16×3×224×224，网络在 RTX 3080 Ti GPU 上进行处理。

相较于其他方法，本文所提方法在 MACs 和 Params 方面具有显著优势。虽然所提方法在参数量方面与 Feng 的方法[19]相似，但其在计算复杂度方面表现出更高的效率。而与其他基于图卷积的方法相对比，本文所提方法不论在计算参数量还是在计算复杂度方面均有明显的优势。

4.4 消融实验

本文在 UCSD Ped2 数据集上进行消融实验以验证自适应时空融合图模块和改进的时空图卷积模块。基础网络为时空图卷积网络，输出帧级别的标签预测概率。

实验结果如表 6 所示，其中时间依赖图仅考虑时间上的连续关系，空间相似图仅考虑了视频中各个片段在空间特征上的相似性关系。实验结果表明，在基础网络架构下，仅考虑单一构图学习异常事件的特征，或对两种构图的结果取平均值，在帧级别 AUC 与准确率上均低于自适应时空融合，表明自适应时空融合能够更好地学习异常事件的时空特征，有利于异常事件的准确定位与准确检测。再加上改进后的时空图卷积模块，帧级别 AUC 和准确率进一步提升了 8.5% 和 9.6%。

此外，在真实世界中视频数据往往会受到各种噪声的影响，这些噪声可能会干扰到异常检测算法的性能，导致误报或漏报。因此，本文进行了相关实验来评估自适应时空融合图模块和改进的时空图卷积模块对噪声的鲁棒性。本文在训练集中引入了随机选择的不同数量的异常样本作为噪声，这些样本来自其他未使用过的异常数据。图 7 清晰地展示了改进后的时空图卷积模块、平均融合的图融合方式 (称为方法 1)、基于 GCN 模块以及自适应时空融合的图融合方式 (称为方法 2) 以及本文提出的方法，这三种情况对噪声鲁棒性的实验结果。

在大多数异常检测设置中，5% 的事件发生率被认为是非常频繁的[41]，在使用这种数量的干扰进行训

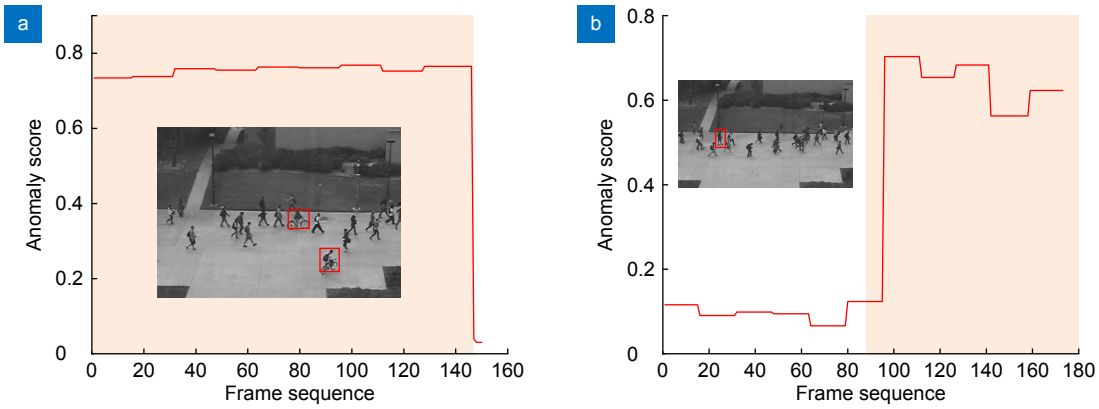


图 4 UCSD Ped2 数据集测试结果。 (a) Test003; (b) Test012

Fig. 4 Test results of UCSD Ped2 dataset. (a) Test003; (b) Test012

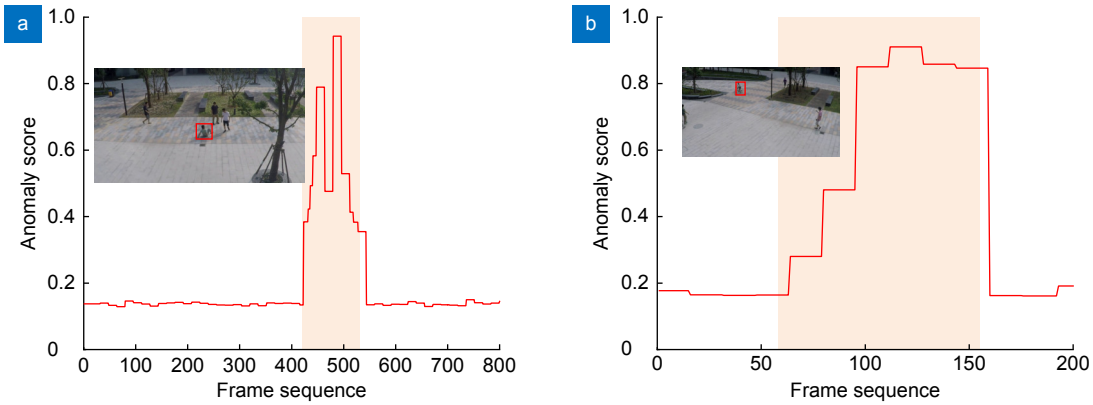


图 5 ShanghaiTech 数据集测试结果。 (a) 04_0004; (b) 12_0173

Fig. 5 Test results of ShanghaiTech dataset. (a) 04_0004; (b) 12_0173

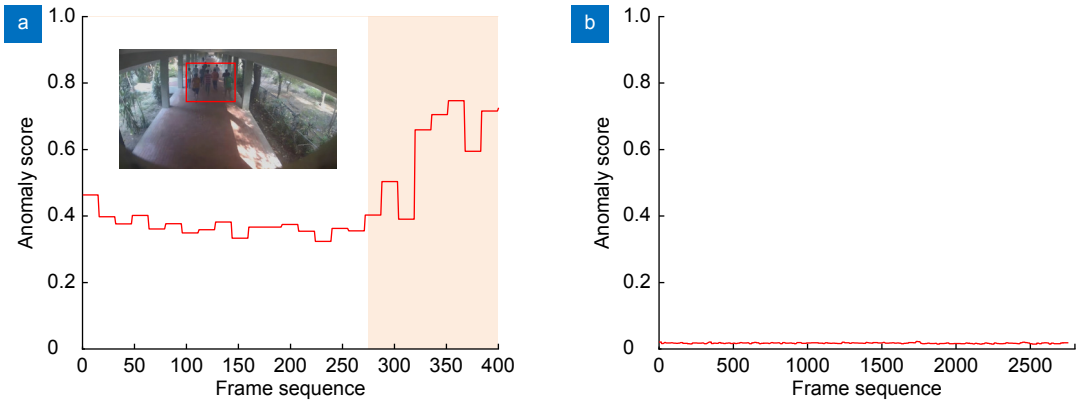


图 6 IITB-Corridor 数据集测试结果。 (a) Test000228; (b) Train000139 (Normal)

Fig. 6 Test results of IITB-Corridor dataset. (a) Test000228; (b) Train000139 (Normal)

表 5 算法复杂度对比

Table 5 Comparison results of different methods on complexity

分类	对比方法	MACs/G	Params/M
基于其他框架的方法	Sultani的方法[7]	154.22	63.33
	Feng的方法[19]	156.86	34.75
基于图卷积的方法	GCN-Anomaly[27]	154.22	63.38
	Chen的方法[33]	154.23	63.90
	本文方法	**109.14**	**19.90**

表 6 消融实验结果

Table 6 Results of ablation experiments

时间依赖图	空间相似图	图融合方式	CRF	AUC/%	准确率/%
√		-		96.6	96.2
	√	-		97.1	96.1
√	√	平均融合[29]		89.2	86.9
√	√	自适应时空融合		96.1	94.2
√	√	**自适应时空融合**	√	**97.7**	**96.5**

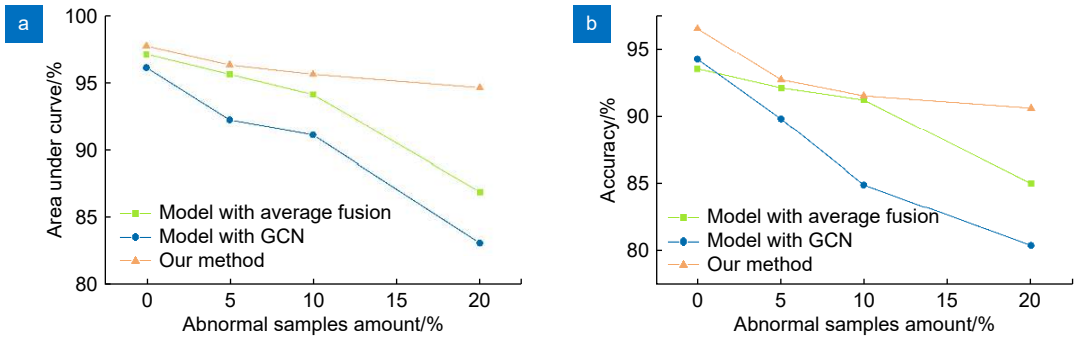


图 7 加噪实验。(a) 使用加噪数据训练的 AUC 损失；(b) 使用加噪数据训练的 ACC 损失

Fig. 7 Noised experiments. (a) AUC loss for training with noise-added data; (b) ACC loss for training with noise-added

练时，三种方法的帧级别 AUC 与准确率的平均性能损失不到 10%。然而，当使用 20% 的异常噪声进行训练时，平均性能损失呈现较大的差异。具体来说，方法 1 与方法 2 的帧级别 AUC 分别损失约 13.1% 与 10.5%；准确率分别损失约 14.5% 与 9%；而本文所提方法在帧级别 AUC 和准确率方面损失仅 3.1% 与 6.1%。因此本文所提方法对于噪声具有一定的鲁棒性，性能损失较小。

综上所述，自适应时空融合模块和改进的时空图卷积模块能够提高异常检测网络的准确性，并提高对噪声的鲁棒性。

5 结 论

为了准确捕捉异常事件中对象的时空相关性，本文提出一种改进时空图卷积网络的视频异常检测方法。在 GCN 模型中引入 CRF 层，对跨帧的时空特征相互作用进行建模，捕捉上下文关系，提高检测准确性。通过自适应融合空间相似图和时间依赖图，在考虑片段间的时空联系的基础上为视频特征学习提供全面信息。在 UCSD Ped2、ShanghaiTech 和 IITB-Corridor 等数据集上进行了实验，证明该方法在帧级别 AUC 和准确率等性能指标上表现良好，可有效用于视频异常事件检测。下一步将以语义方式同时考虑前景对象和背景信息以实现更精确的检测，并进一步研究模型的泛化性能，以适应未知场景下的异常事件检测。

参考文献

[1] Gong Y L, Zhang X X, Chen S. Survey on deep learning approach for video anomaly detection[J]. *Data Commun*, 2023(3): 45−49.
龚益玲, 张鑫昕, 陈松. 基于深度学习的视频异常检测研究综述[J]. 数据通信, 2023(3): 45−49.

[2] Wang X G, Yan Y L, Tang P, et al. Revisiting multiple instance neural networks[J]. *Pattern Recognit*, 2018, **74**: 15−24.

[3] Zhou Z H, Sun Y Y, Li Y F. Multi-instance learning by treating instances as non-I. I. D. samples[C]//*Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, 2009: 1249–1256. https://doi.org/10.1145/1553374.1553534.

[4] Cheng W, Chen Z B, Li Q Q, et al. Multiple object tracking with aligned spatial-temporal feature[J]. *Opto-Electron Eng*, 2023, **50**(6): 230009.
程稳, 陈忠碧, 李庆庆, 等. 时空特征对齐的多目标跟踪算法[J]. 光电工程, 2023, **50**(6): 230009.

[5] Li J, Liu Y, Zou L. A dynamic graph convolutional network based on spatial-temporal modeling[J]. *Acta Sci Nat Univ Pekins*, 2021, **57**(4): 605−613.
李荆, 刘钰, 邹磊. 基于时空建模的动态图卷积神经网络[J]. 北京大学学报 (自然科学版), 2021, **57**(4): 605−613.

[6] Lv J, Wang Z Y, Liang H C. Boundary attention assisted dynamic graph convolution for retinal vascular segmentation[J]. *Opto-Electron Eng*, 2023, **50**(1): 220116.
吕佳, 王泽宇, 梁浩城. 边界注意力辅助的动态图卷积视网膜血管分割[J]. 光电工程, 2023, **50**(1): 220116.

[7] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos[C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018: 6479–6488. https://doi.org/10.1109/CVPR.2018.00678.

[8] Zhang J G, Qing L Y, Miao J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection[C]//*2019 IEEE International Conference on Image Processing* (*ICIP*), Taipei, China, 2019: 4030–4034. https://doi.org/10.1109/ICIP.2019.8803657.

[9] Li S, Liu F, Jiao L C. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection[C]//*Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022: 1395–1403. https://doi.org/10.1609/aaai.v36i2.20028.

[10] Liang W J, Zhang J M, Zhan Y Z. Weakly supervised video anomaly detection based on spatial–temporal feature fusion enhancement[J]. *Signal, Image Video Process*, 2024, **18**(2): 1111−1118.

[11] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]//*Proceedings of the 5th International Conference on Learning Representations*, Toulon, 2017.

[12] Zhou H, Zhan Y Z, Mao Q R. Video anomaly detection based on space-time fusion graph network learning[J]. *J Comput Res Dev*, 2021, **58**(1): 48−59.
周航, 詹永照, 毛启容. 基于时空融合图网络学习的视频异常事件

检测[J]. 计算机研究与发展, 2021, **58**(1): 48−59.

[13] Purwanto D, Chen Y T, Fang W H. Dance with self-attention: a new look of conditional random fields on anomaly detection in videos[C]//*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, 2021: 173–183. https://doi.org/10.1109/ICCV48922.2021.00024.

[14] Mu H Y, Sun R Z, Wang M, et al. Spatio-temporal graph-based CNNs for anomaly detection in weakly-labeled videos[J]. *Inf Process Manage*, 2022, **59**(4): 102983.

[15] Liu M T, Li X R, Liu Y G, et al. Weakly supervised anomaly detection with multi-level contextual modeling[J]. *Multimedia Syst*, 2023, **29**(4): 2153−2164.

[16] Cheng K, Zeng X H, Liu Y, et al. Spatial-temporal graph convolutional network boosted flow-frame prediction for video anomaly detection[C]//*ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, 2023: 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095170.

[17] Li X B, Wang W, Li Q Y, et al. Spatial-temporal graph-guided global attention network for video-based person re-identification[J]. *Mach Vision Appl*, 2024, **35**(1): 8.

[18] Wan B Y, Fang Y M, Xia X, et al. Weakly supervised video anomaly detection via center-guided discriminative learning[C]//*2020 IEEE International Conference on Multimedia and Expo (ICME)*, London, 2020: 1–6. https://doi.org/10.1109/ICME46284.2020.9102722.

[19] Feng J C, Hong F T, Zheng W S. MIST: multiple instance self-training framework for video anomaly detection[C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, 2021: 14004–14013. https://doi.org/10.1109/CVPR46437.2021.01379.

[20] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//*Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, 2001: 282–289.

[21] Gao H C, Pei J, Huang H. Conditional random field enhanced graph convolutional neural networks[C]//*Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 2019: 276–284. https://doi.org/10.1145/3292500.3330888.

[22] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C]*//Proceedings of the 24th International Conference on Neural Information Processing Systems*, Granada, 2011: 109–117.

[23] Zhang J W, Zhang X L, Zhu Z Q, et al. Efficient combination graph model based on conditional random field for online multi-object tracking[J]. *Complex Intell Syst*, 2023, **9**(3): 3261−3276.

[24] Chen D Y, Wang P T, Yue L Y, et al. Anomaly detection in surveillance video based on bidirectional prediction[J]. *Image Vision Comput*, 2020, **98**: 103915.

[25] Lu C W, Shi J P, Jia J Y. Abnormal event detection at 150 FPS in MATLAB[C]//*Proceedings of the 2013 IEEE International Conference on Computer Vision*, Sydney, 2013: 2720–2727. https://doi.org/10.1109/ICCV.2013.338.

[26] Rodrigues R, Bhargava N, Velmurugan R, et al. Multi-timescale trajectory prediction for abnormal human activity detection[C]//*Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision*, Snowmass, 2020: 2615–2623. https://doi.org/10.1109/WACV45572.2020.9093633.

[27] Zhong J X, Li N N, Kong W J, et al. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection[C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, 2019: 1237–1246. https://doi.org/10.1109/CVPR.2019.00133.

[28] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016: 733–742. https://doi.org/10.1109/CVPR.2016.86.

[29] Gong D, Liu L Q, Le V, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, 2019: 1705–1714. https://doi.org/10.1109/ICCV.2019.00179.

[30] Yu G, Wang S Q, Cai Z P, 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*, Seattle, 2020: 583–591. https://doi.org/10.1145/3394171.3413973.

[31] Taghinezhad N, Yazdi M. A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction[J]. *IEEE Access*, 2023, **11**: 9295−9310.

[32] Tian Y, Pang G S, Chen Y H, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]//*Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, 2021: 4955–4966. https://doi.org/10.1109/ICCV48922.2021.00493.

[33] Chen H Y, Mei X, Ma Z Y, et al. Spatial–temporal graph attention network for video anomaly detection[J]. *Image Vision Comput*, 2023, **131**: 104629.

[34] Wang L, Tian J W, Zhou S P, et al. Memory-augmented appearance-motion network for video anomaly detection[J]. *Pattern Recognit*, 2023, **138**: 109335.

[35] Tur A O, Dall 'Asen N, Beyan C, et al. Exploring diffusion models for unsupervised video anomaly detection[C]//*2023 IEEE International Conference on Image Processing (ICIP)*, Kuala Lumpur, 2023: 2540–2544. https://doi.org/10.1109/ICIP49359.2023.10222594.

[36] Acsintoae A, Florescu A, Georgescu M I, et al. UBnormal: new benchmark for supervised open-set video anomaly detection[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 2022: 20111–20121. https://doi.org/10.1109/CVPR52688.2022.01951.

[37] Zeng X L, Jiang Y L, Ding W R, et al. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos[J]. *IEEE Trans Circuits Syst Video Technol*, 2023, **33**(1): 200−212.

[38] Li J, Huang Q W, Du Y J, et al. Variational abnormal behavior detection with motion consistency[J]. *IEEE Trans Image Process*, 2022, **31**: 275−286.

[39] Cao C Q, Lu Y, Wang P, et al. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation[C]//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 2023: 20392–20401. https://doi.org/10.1109/CVPR52729.2023.01953.

[40] Majhi S, Dai R, Kong Q, et al. Human-Scene Network: a novel baseline with self-rectifying loss for weakly supervised video

anomaly detection[J]. *Comput Vis Image Underst*, 2024, **241**: 103955.

[41] Markovitz A, Sharir G, Friedman I, et al. Graph embedded pose clustering for anomaly detection[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 2020: 10536–10544. https://doi.org/10.1109/CVPR42600.2020.01055.

作者简介

【通信作者】张红民 (1970-)，男，博士，教授，主要研究方向为图像处理与模式识别。

E-mail：hmzhang@cqut.edu.cn

田钱前 (1999-)，女，硕士研究生，主要研究方向为图像处理、深度学习。

E-mail：qianqiantian@stu.cqut.edu.cn

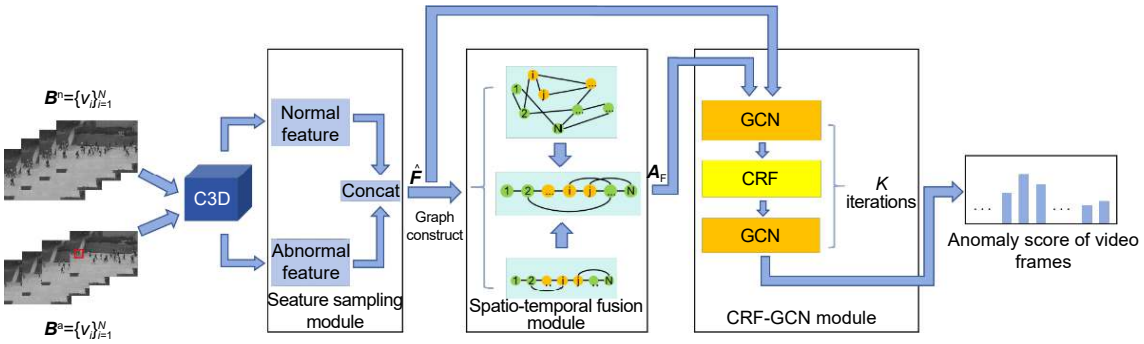
颜鼎鼎 (2000-)，女，硕士研究生，主要研究方向为图像处理、计算机视觉。

E-mail：ydd0010@stu.cqut.edu.cn

扫描二维码，获取PDF全文

Improved spatio-temporal graph convolutional networks for video anomaly detection

Zhang Hongmin*, Yan Dingding, Tian Qianqian



Improved spatio-temporal graph convolutional network model framework

Overview: Video surveillance systems are increasingly widely used in public places and play an important role in maintaining social security and stability. However, the collection and labeling of anomalous videos are subject to subjective factors, resulting in video data containing only video-level labels and lacking detailed information, limiting the intelligent analysis of videos, especially in the field of anomaly detection, where richer data information is needed to improve model performance.

Video data is typical spatio-temporal data, the spatio-temporal features shown by the abnormal events in the video have significant correlation, and the connection between the segments in the video can be constructed by introducing the graph structure in both time perspective and space perspective, but the traditional convolution operation can not be directly applied to the graph. Although Graph Convolutional Neural Network (GCN) can effectively process data with the graph structure, it is still deficient in capturing the intrinsic relationship between objects in neighbouring frames, especially in coping with the complex spatio-temporal dependencies between frames in a video sequence. To model the spatio-temporal correlations of video segments more reasonably under the graph structure, and then effectively detect and locate video anomalies, this paper proposes an improved video anomaly detection method with spatio-temporal graph convolutional networks. Each clip in the video is regarded as a node; two key graph models, a spatial similarity graph, and a temporal dependency graph are constructed. The video features are learned by adaptive fusion based on the consideration of spatio-temporal connections between clips. Since anomalous events can be formed through spatio-temporal interactions between multiple objects, taking advantage of the good graph modeling benefits of Conditional Random Field (CRF), a CRF layer is introduced into the GCN model to model the interactions between spatio-temporal features across frames to capture their contextual relationships, thus improving the detection accuracy of the model.

Experiments were conducted on three video anomaly event datasets, including UCSD Ped2, ShanghaiTech, and IITB-Corridor. The frame-level AUC values reach 97.7%, 90.4%, and 86.0%, respectively, and the experimental results verify the effectiveness of the proposed method.

Zhang H M, Yan D D, Tian Q Q. Improved spatio-temporal graph convolutional networks for video anomaly detection[J]. *Opto-Electron Eng*, 2024, **51**(5): 240034; DOI: 10.12086/oee.2024.240034

Foundation item: Project supported by the National Natural Science Foundation of China (61901068), Chongqing Natural Science Foundation Top Project (cstc2021 jcyj-msxmX0525, CSTB2022NSCQ-MSX0786, CSTB2023NSCQ-MSX0911), and Science and Technology Research Project of Chongqing Municipal Education Commission (KJQN202201109)

School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China

* E-mail: hmzhang@cqut.edu.cn