E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2024)07-1998-13

论文引用格式:Zhu X R, Qian X Y, Shi Y Z, Tao X D and Li Z Y. 2024. Video anomaly detection with long-and-short-term time series correlations. Journal of Image and Graphics, 29(07):1998-2010(朱新瑞, 钱小燕, 施俞洲, 陶旭东, 李智昱. 2024. 长短期时间序列关联的视频异常事件检测. 中国图象图形学报, 29(07):1998-2010)[DOI:10.11834/jig.230406]

长短期时间序列关联的视频异常事件检测

朱新瑞,钱小燕*,施俞洲,陶旭东,李智昱

南京航空航天大学民航学院,南京　211106

摘　要:**目的**　多示例学习是解决弱监督视频异常事件检测问题的有力工具。异常事件发生往往具有稀疏性、突发性以及局部连续性等特点,然而,目前的多示例学习方法没有充分考虑示例之间的联系,忽略了视频片段之间的时间关联,无法充分分离正常片段和异常片段。针对这一问题,提出了一种长短期时间序列关联的二阶段异常检测网络。**方法**　第1阶段是长短期时间序列关联的异常检测网络(long-and-short-term correlated mil abnormal detection framework, LSC-transMIL),将Transformer结构应用到多示例学习方法中,添加局部和全局时间注意力机制,在学习不同视频片段间的空间关联语义信息的同时强化连续视频片段的时间序列关联;第2阶段构建了一个基于时空注意力机制的异常检测网络,将第1阶段生成的异常分数作为细粒度伪标签,使用伪标签训练策略训练异常事件检测网络,并微调骨干网络,提高异常事件检测网络的自适应性。**结果**　实验在两个大型公开数据集上与同类方法比较,两阶段的异常检测模型在UCF-crime、ShanghaiTech数据集上曲线下面积(area under curve, AUC)分别达到82.88%和96.34%,相比同为两阶段的方法分别提高了1.58%和0.58%。消融实验表明了关注时间序列的Transformer模块以及长短期注意力的有效性。**结论**　本文将Transformer应用于时间序列的多示例学习,并添加长短期注意力,突出局部异常事件和正常事件的区别,有效检测视频中的异常事件。

关键词:异常检测;Transformer网络;时空注意力;多示例学习(MIL);弱监督

Video anomaly detection with long-and-short-term time series correlations

Zhu Xinrui, Qian Xiaoyan*, Shi Yuzhou, Tao Xudong, Li Zhiyu

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing　211106, China

Abstract: **Objective** Video anomaly detection has been applied in many fields such as manufacturing, traffic management and security monitoring. However, detailed annotation of video data is labor intensive and cumbersome. Consequently, many researchers have started to employ weakly supervised learning methods to address this issue. Unlike the supervised learning method, the weakly supervised learning only requires video-level labels in the training stage, which greatly reduces the workload of dataset labeling, and only frame-level labeling information is required for the test dataset. Multiple instance learning (MIL) has been recognized as a powerful tool for addressing weakly supervised video abnormal event detection. Abnormal behavior in video is highly correlated with video context information. The traditional MIL method uses convolutional 3D network to extract video features, uses the ordering loss function, and introduces sparsity and time smoothing constraints into the ordering loss function to integrate time information into the ordering model. Introducing time concern only into the loss function is not enough. The use of temporal convolutional network to extract video context infor-

收稿日期:2023-06-28;**修回日期**:2023-09-28;**预印本日期**:2023-10-04

*通信作者:钱小燕　qianxiaoyan@nuaa.edu.cn

基金项目:国家自然科学基金项目(61803199,U2033201)

Supported by: National Natural Science Foundation of China (61803199, U2033201)

mation further enhances the effect of video anomaly detection network. However, this global introduction of time information cannot sufficiently separate abnormal video clips from normal video clips. Therefore, the attention MIL builds time-enhancing networks to learn motion features while using the attention mechanism to incorporate temporal information into the ranking model. The learned attention weights can help better distinguish between abnormal and normal video clips. The spatiotemporal fusion graph network constructs spatial similarity graphs and temporal continuity graphs separately for video segments, which are then fused to generate a spatiotemporal fusion graph. This approach strengthens the spatiotemporal correlations among video segments, ultimately enhancing the accuracy of abnormal behavior detection. Multiple instance self-training framework uses pseudo-label training, which is an effective training strategy to improve model quality in weakly supervised learning. It constructs a two-stage training network and uses the pseudo-label trained by the first-stage MIL to guide the training of the second-stage self-guided attention feature extractor, providing a general idea to improve model quality. However, these approaches do not fully exploit temporal correlations, as the feature representation of the instances lacks fusion with neighboring and global features. Abnormal events often exhibit characteristics such as sparsity, suddenness, and local continuity, and the insufficient temporal correlations between video segments result in an inadequate separation between normal and abnormal segments. To address this issue, this paper proposes a two-stage abnormal detection network with long-and-short-term time series association. **Method** The first stage involves a long-and-short-term time series association abnormal detection network (LSC-transMIL) that applies the Transformer structure to MIL methods. It consists of two layers, each containing a local temporal sequence correlation attention module and a global instance correlation attention module. The former learns information in the temporal dimension between individual instances and neighboring instances, while the latter focuses on the association between individual instances and global information. Combining local and global attention mechanisms makes it possible to establish meaningful information correlations among instances, highlighting the distinctions between local and global features in the video. This approach makes it easier to distinguish abnormal video segments from normal ones. This module generates new instance features, which are then fed into the ranking model to generate video abnormal scores and pseudo-labels. In the second stage, a spatiotemporal attention mechanism-based abnormal detection network is constructed. The SlowFast backbone network is employed to extract video features, and the slow and fast pathway features are weighted and fused using spatiotemporal attention. The slow branch pays attention to the spatiotemporal information of the video frame using the spatiotemporal attention module, while the fast branch guides the attention to the temporal information through the time-dimensional attention module, and then the two branch features are spliced to obtain the final video features. The abnormal scores generated in the first stage are used as fine-grained pseudo-labels to train the abnormal event detection network by using a pseudo-labeling strategy. Furthermore, the backbone network is fine-tuned to enhance the adaptive capability of the abnormal event detection network. **Result** Extensive experiments were conducted on two large-scale public datasets (UCF-crime and ShanghaiTech) to compare the proposed two-stage abnormal detection model with similar methods. The two-stage model achieved area under the curve scores of 82.88% and 96.34% on the UCF-crime and ShanghaiTech datasets, respectively, demonstrating an improvement of 1.58% and 0.58% compared with other two-stage methods. Sufficient ablation experiments were conducted on the two datasets, and the effects of the proposed LSC-transMIL, traditional MIL method, and attention MIL method were compared under three backbone networks, proving the effectiveness of LSC-transMIL. Qualitative and quantitative explanations are given for the ablation experiments of global attention and global local attention, and the effectiveness of combining local and global attention is proved. The role of local and global time correlation is visualized using heat maps. **Conclusion** This paper applies the Transformer to time series-based MIL and introduces long-and-short-term attention to highlight the differences between local abnormal events and normal events. The proposed two-stage abnormal detection network utilizes the abnormal scores generated in the first stage as pseudo-labels, trains a network based on the SlowFast backbone network and spatiotemporal attention modules, and fine-tunes the backbone network to enhance the adaptive capability of the abnormal detection network. The proposed approach effectively improves the accuracy of abnormal event detection.

Key words: anomaly detection; Transformer; spatio-temporal attention; multiple instance learning(MIL); weakly supervised

0 引 言

视频异常检测任务旨在自动检测和定位异常事件，如交通事故和犯罪行为。由于背景复杂、数据维数高和异常定义不清晰等难点，视频异常检测一直是计算机视觉领域的一项艰巨任务。在现实场景中，异常事件种类繁多，且远少于正常事件，因此要收集各种异常进行建模几乎是不可能的（梁家菲等，2023）。此外，获取精确的帧级标签既耗时又费力，所以通常使用基于多示例学习的弱监督方法进行视频异常检测（Abbas和Al-Ani，2022）。

弱监督视频异常检测通常使用动作识别领域预训练模型提取视频特征，然后使用多示例学习（multiple instance learning，MIL）对视频特征进行异常分数排序，最终得到视频片段发生异常事件的概率。Sultani等人（2018）首次将MIL方法引入视频异常检测任务中，使用多示例排序框架获得视频异常分数（如图1（a）所示），在排序损失函数中引入稀疏性和时间平滑度约束，以在训练期间更好地定位异常。Zhu和Newsam（2019）通过注意力机制和时序排序损失将时间上下文纳入了MIL排序模型，学习到的注意力权重有助于更好地区分异常视频片段和正常视频片段，如图1（b）所示。更进一步地，Feng等人（2021）提出一种有效的伪标签训练策略，使用MIL方法生成视频异常分数作为视频片段的伪标签，依据这一更为精细的片段标签构建分类网络有效提高了异常检测准确率。然而，现有方法构建示例特征时将视频片段视为独立同分布，没有充分考虑示例之间的信息交互（Shao等，2021）。Sultani等人（2018）仅在排序损失中引入稀疏性和时间平滑度约束，Zhu和Newsam（2019）在排序模型中添加时间上下文信息，而每个示例特征中未充分构建与其他示例的信息融合，在视频异常检测任务中，忽略了视频片段间的相关性，对于联系较强的事件易出现漏检。

Transformer由于具有很强的描述序列中不同片段之间的相关性以及建模远程信息的能力，广泛应用于许多视觉任务，如目标识别与跟踪、医学图像分类等（Shao等，2021）。因此可以借鉴Transformer结构构建多示例学习网络，加强示例特征之间的时空联系。针对异常事件的稀疏性、突发性以及具有局部连续性等特点，为了避免异常事件的特征淹没在正常事件特征中、网络对于异常事件不敏感等问题，本文提出了一个基于长短期时间序列关联的MIL二阶段异常检测网络（two-stages long-and-short-term correlated MIL abnormal detection framework，LSC-transMIL-stageT）。首先引入多示例Transformer模型去学习各个片段之间的相关性。构建全局注意力机制学习各视频片段之间的内在相关信息；并通过局部注意力机制对连续视频片段的局部信息进行建模，学习局部片段与临近片段的时间序列上的信息变化，对局部和全局相关信息进行自适应融合，以期提高视频异常事件检测的准确性。在此基础上，采用伪标签异常训练策略，构建二阶段基于SlowFast骨干网络的时空注意力模型，进一步增强异常检测性能。第1阶段基于LSC-transMIL方法为每一视频片段生成高可信度的细粒度异常分数，以此作为伪标签引导二阶段的网络训练。训练过程中微调主干网络和时空注意力模块，可以消除在行为识别训练集上预训练模型的偏向，增强对异常检测的敏感性，提高异常检测性能。

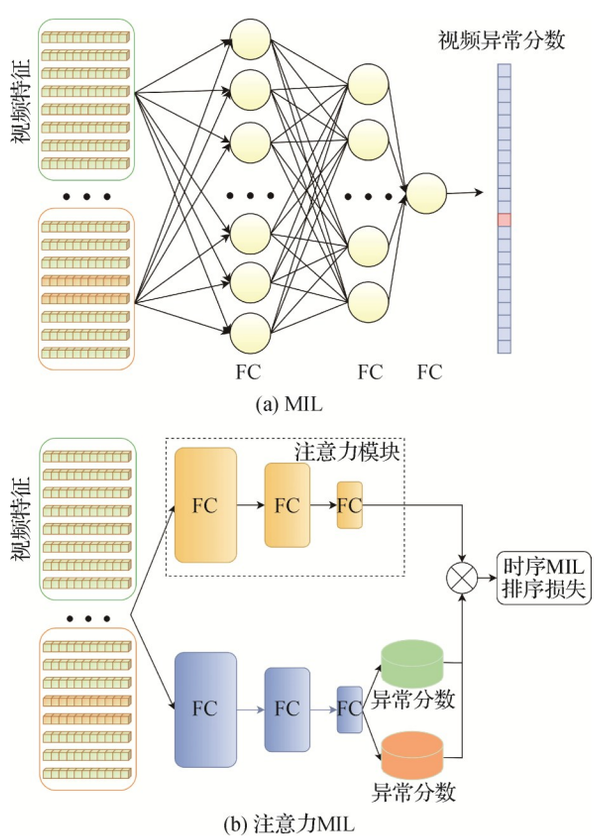


图1 MIL和注意力MIL

Fig. 1 MIL and attention MIL ((a) MIL; (b) attention MIL)

本文主要工作如下：1）针对现有基于MIL的异常检测方法示例间联系不足的问题，提出了基于Transformer的LSC-transMIL，将Transformer模型应用到MIL异常检测中，并加入局部和全局时间注意力机制强化示例间的时间相关性，从而提高异常判别的性能，并根据此模型生成视频片段伪标签。2）为增强异常检测性能，建立了二阶段的异常检测时空注意力模型。将第1阶段LSC-transMIL生成的伪标签用于引导第2阶段骨干网络的训练；并设计时空注意力模型加强不同示例间的关联；训练过程中微调模型参数，强化异常检测的自适应能力。3）对MIL常用方法以及现有经典的弱监督异常检测方法进行了比较分析。实验表明，本文提出的基于LSC-transMIL-stageT的两阶段视频异常检测网络，在上海科技大学数据集上曲线下面积（area under curve，AUC）达到96.22%，在UCF-crime数据集上AUC达到82.88%，具有很好的异常检测性能。

1 相关工作

1.1 弱监督视频异常检测

视频异常检测依据标签信息的详细程度可分为监督学习、弱监督学习和无监督学习3类，其中弱监督学习在训练时仅需要视频级别的标签，测试时需要帧级别的标签，相对来说数据集标注难度较低，同时异常识别准确率也较高，是实现视频异常检测的一种主流方法（王志国和章毓晋，2020）。

2018年，Sultani等人（2018）首次将多示例学习方法应用到视频异常检测领域，将视频片段视为多示例学习中的示例，使用C3D（convolutional 3D）网络提取视频特征，并建立排序模型预测视频片段的异常分数。在此基础上，Zhu和Newsam（2019）通过注意力机制引入MIL排序模型以提高异常检测准确率，该方法将注意力应用在异常分数生成阶段，而不是视频特征生成阶段，不能达到最优检测效果。Zhang等人（2019）在提取视频特征时添加时间卷积网络（temporal convolutional network，TCN）来提取视频上下文信息使得异常检测性能得到了提升。但该方法仅获取长时间相关性，而突发性的短暂异常事件特征可能会被正常特征所掩盖。于是Majhi（2020）在使用3D卷积神经网络（3D convolutional neural network，3DCNN）提取特征的同时，通过两级时间注意力网络增强视频特征，实现异常事件的检测和分类。周航等人（2021）尝试从时空融合图网络角度解决这一问题，对视频片段的特征分别构建空间相似图和时间连续图，从而生成时空融合图，加强了视频片段时空关联，提高异常检测准确率。Feng等人（2021）更进一步提出一种两阶段的伪标签训练策略。第1阶段使用MLP网络生成视频异常分数作为伪标签；第2阶段构建基于伪标签的异常分类网络，可有效提高异常检测效果。但是这些方法将包（bag）内视频片段视为相互独立的，在特征构建时没有充分考虑示例间的关联，本文将Transformer结构应用于MIL方法中，以加强示例之间的联系。

1.2 MIL相关工作

多示例学习（MIL）与传统单示例学习、多标签学习不同，MIL的单个样本称为包（bag），每个包中含有多个示例，示例在训练过程没有标签，仅有包级别的标签（Carbonneau等，2018）。多示例学习可分为嵌入方法、示例方法以及包方法（Wei和Zhou，2016），由于视频异常检测需要获取每个视频帧或者视频片段的异常分数，因此往往使用基于示例的方法。

常规MIL方法通常基于示例之间是独立同分布的这一假设（Shao等，2021），但是这样会忽略示例之间的联系这一有效信息。Ilse等人（2018）基于门控注意力机制提出注意力池化方法，改进常规多示例学习方法所使用的最大值池化和平均值池化不能通过训练优化的问题；而通过注意力机制学习示例权重，然后连接全连接层获得示例预测标签，其准确度很大程度上取决于学习到的权重优劣，难以直接优化。于是Shi等人（2020）将这两个过程进行结合，提出了一种全新的基于注意力机制的损失函数，在计算权重的同时完成预测。由于Transformer具有很强的描述序列中不同片段之间的相关性以及建模远程信息的能力，Shao等人（2021）构建了transMIL（Transformer based MIL）结构，用于加强每个示例（局部图像）之间的上下文联系，在全视野数字切片（whole slide image）领域表现出色。Li等人（2022）提出了多序列学习（multi-sequence learning，MSL）方法，不再使用示例作为优化单元，而是使用由多个示例组成的序列作为优化单元，从而减少了训练初期模型错误选择异常示例的可能性。因此可以借鉴Transformer的深度语义学习能力、长序列建模能力，构建异常视频的内在全局和局部的相关性，更好地

实现视频异常事件的检测。

2 本文方法

2.1 视频异常检测问题描述

多示例学习的视频异常检测方法描述如下：定义输入为 B 个包（即视频）$\boldsymbol{M}_i$ 的集合 $\{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_B\}$。其中每个包包含 t 个视频片段，即 t 个示例 $\boldsymbol{M}_i = \{\boldsymbol{m}_{i,1}, \boldsymbol{m}_{i,2}, \cdots, \boldsymbol{m}_{i,j}, \cdots, \boldsymbol{m}_{i,t}\}$，$i = 1, 2, \cdots, B$，$j = 1, 2, \cdots, t$，每个包中片段数目并不相同，实际运算中为保证维度统一选择片段数目最大值记为 T。每个包的标签记为 $L_i \in \{0, 1\}$，0 表示正常视频，1 表示视频中包含异常事件。视频异常检测的目的是生成视频包 $\boldsymbol{M}_i$ 中每个示例 $\boldsymbol{m}_{i,j}$ 的异常分数，即 $\boldsymbol{S}_i = \{s_{i,1}, s_{i,2}, \cdots, s_{i,j}, \cdots, s_{i,t}\}$，$s_{i,j} \in [0, 1]$，并根据异常分数判断当前视频片段是否发生异常。

传统的 MIL 方法将视频片段看做相互独立的，如图 1（a）所示。使用多示例排序模型获得视频片段的异常分数，排序模型可描述为包含异常事件的包中示例对应的异常分数最大值大于正常视频包中示例对应的异常分数最大值（Sultani 等，2018），即

$$\max_{i \in \boldsymbol{P}_a} f\left(\boldsymbol{m}_{ij}^a\right) > \max_{i \in \boldsymbol{P}_n} f\left(\boldsymbol{m}_{ij}^n\right) \tag{1}$$

式中，$f\left(\boldsymbol{m}_{i,j}\right)$ 代表视频片段对应的异常分数，$\boldsymbol{P}_a$，$\boldsymbol{P}_n$ 分别代表异常视频和正常视频，$\max_{i \in \boldsymbol{P}_a}(\cdot)$ 代表异常视频中示例异常分数最大值，$\max_{i \in \boldsymbol{P}_n}(\cdot)$ 代表正常视频中示例异常分数最大值。

注意力 MIL 方法如图 1（b）所示，为各示例片段增加了注意力权重（Zhu 和 Newsam，2019），具体为

$$\max_{i \in \boldsymbol{P}_a} w_i f\left(\boldsymbol{m}_{ij}^a\right) > \max_{i \in \boldsymbol{P}_n} w_i f\left(\boldsymbol{m}_{ij}^n\right) \tag{2}$$

式中，w_i 表示习得注意力权重。正如以上模型所示，这两种 MIL 方法将示例特征视为相互独立的，某一片段特征未充分关联包内其他片段信息。而 Shao 等人（2021）验证了相关示例间的信息熵小于独立示例的信息熵，即

$$H_i\left(\boldsymbol{m}_{i,1}, \boldsymbol{m}_{i,2}, \cdots, \boldsymbol{m}_{i,j}\right) = \sum_{j=2}^{t} H\left(\boldsymbol{m}_{i,t} \middle| \boldsymbol{m}_{i,1}, \boldsymbol{m}_{i,2}, \cdots, \boldsymbol{m}_{i,j-1}\right) + H\left(\boldsymbol{m}_{i,1}\right) \leqslant \sum_{j=1}^{t} H\left(\boldsymbol{m}_{i,j}\right) \tag{3}$$

式中，$H_i\left(\boldsymbol{m}_{i,1}, \boldsymbol{m}_{i,2} \cdots, \boldsymbol{m}_{i,j}\right)$ 表示对于包 $\boldsymbol{M}_i$ 构建示例之间联系后的包的信息熵，$\sum_{j=1}^{t} H\left(\boldsymbol{m}_{i,j}\right)$ 表示独立同分布下包 $\boldsymbol{M}_i$ 的信息熵。可见构建示例间的相关性可以有效降低示例的不确定性，从而生成更多对多示例异常检测有用的信息。因此，不同于现有方法，本文基于 Transformer 结构构建示例间的长短期注意力机制，学习不同视频片段间的时空语义相关信息，这样示例的异常判断不仅取决于其本身特征，还取决于相关示例片段的影响，可降低示例信息不确定性，从而提高视频异常检测的准确率。

2.2 一阶段 LSC-transMIL 视频异常检测框架

本文提出的基于长短期相关的 MIL 异常检测框架如图 2 所示，主要分为 3 个部分：1）特征提取，将正常视频和异常视频集合 $\{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_B\}$ 输入骨干网络提取视频段特征，记为 $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_B\}$，$\boldsymbol{X} \in \mathbf{R}^{B \times T \times F}$，$B$ 代表多示例学习中包的数量，T 代表各包中示例的数量最大值，F 为每个示例即视频片段特征的维度。2）视频片段的长短期相关语义学习，所提 LSC-transMIL 模块包含两层 Transformer 结构。网络输入为视频特征 $\boldsymbol{X}_i \in \mathbf{R}^{1 \times T \times F}$，每层学习并融合各视频片段的深层局部和全局相关信息，每层输出记为 $\boldsymbol{Y}_i \in \mathbf{R}^{1 \times T \times F}$，与输入特征维度保持一致。3）异常检测，将包特征集合 $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_B\}$，$\boldsymbol{Y} \in \mathbf{R}^{B \times T \times F}$ 输入多层感知器（multi-layer perception，MLP）网络，对视频片段进行异常检测，输出视频帧异常分数 $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_B\}$，$\boldsymbol{S} \in \mathbf{R}^{B \times T \times 1}$，1 维度表示当前视频片段异常分数。

同时，第 1 阶段生成第 2 阶段需要的伪标签 $\boldsymbol{L}_{P,i} = \{l_{i,1}, l_{i,2}, \cdots, l_{i,j}, \cdots, l_{i,t}\}$，$l_{i,j} \in [0, 1]$，设置阈值为 ψ，伪标签生成计算式为

$$l_{ij} = \begin{cases} s_{ij} & s_{ij} \geqslant \psi \\ 0 & s_{ij} < \psi \end{cases} \tag{4}$$

2.3 LSC-transMIL 模型

本文 LSC-transMIL 模块包含两层结构，每层包括一个局部时间序列相关性注意力模块和全局示例相关性注意力模块，前者学习单个示例与临近示例间时间维度上的语义信息，后者充分利用各示例的空间特征学习不同示例间的空间语义相关性，避免视频异常的稀疏性导致过拟合的现象。

局部时间序列相关性注意力（local time correlation attention）模块使用可训练的高斯核函数建模临近示例之间的相关性，强调局部时间连续性，削弱单

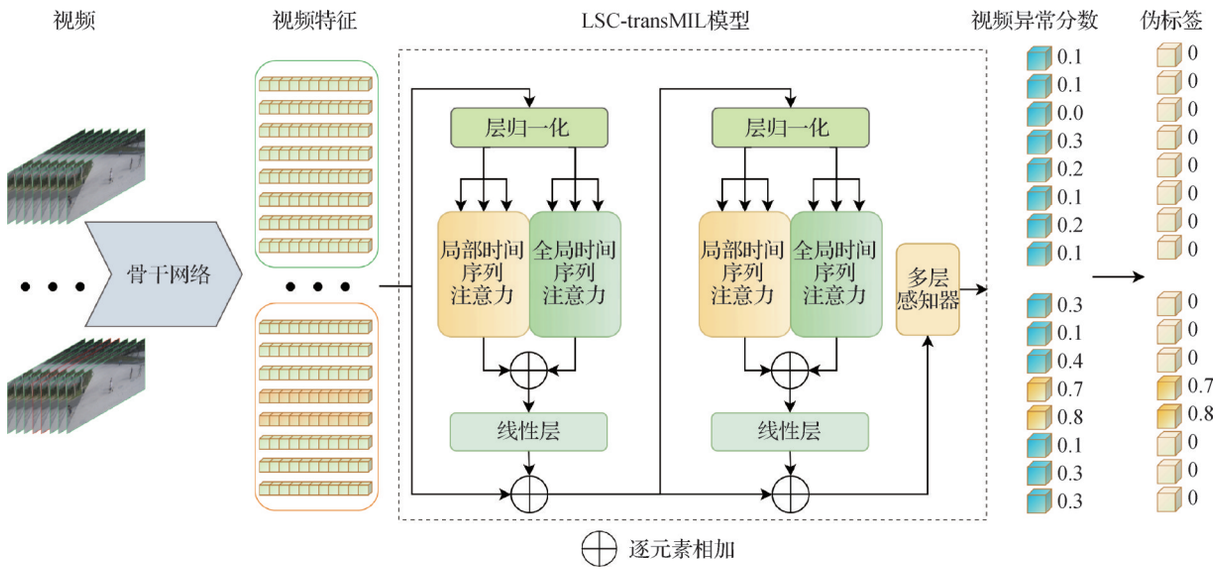


图 2　LSC-transMIL 网络结构

Fig. 2　LSC-transMIL network structure

个示例与全局序列的关联。局部注意力模型为

$$\boldsymbol{\Gamma} = f_{\text{softmax}}\left(\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\left[\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{ij\in\{1,\cdots,T\}}\right) \tag{5}$$

式中，i 表示当前时刻，$|j-i|$ 表示 j 时刻与当前时刻的时间距离，$f_{\text{softmax}}(\cdot)$ 函数对权重进行归一化，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 与常规 Transformer 结构中的含义保持一致。

全局示例相关性注意力（global correlation attention）模块使用 Transformer 方法构建当前示例与包内所有示例之间的联系，具体为

$$\boldsymbol{G} = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{6}$$

由于视频异常检测中异常事件往往具有小概率、短时间的性质，异常事件往往与临近的示例表现出更强的关联性，而和全局信息会有较大差异。如果过于关注全局关联，会将异常事件淹没在正常事件中或者正常示例被误判为异常示例，故而在注意力模块更加关注局部示例之间的关联，通过权重超参量 α 调整全局注意力和局部注意力的异常检测贡献度。输入 $\boldsymbol{X}_i$ 的注意力矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A} = (\alpha\boldsymbol{\Gamma} + \boldsymbol{G})\boldsymbol{V} \tag{7}$$

经过注意力机制及归一化层 L_1 得到 LSC-transMIL 模块的输出，具体为

$$\boldsymbol{Y}_i = L_1(\boldsymbol{A} + \boldsymbol{X}_i),\ \boldsymbol{Y}_i \in \mathbf{R}^{1\times T\times F} \tag{8}$$

最后将获得的新示例特征，输入 MIL 排序模型，本文异常检测 MIL 排序模型可表示为

$$\max_{i\in P_a} f_{j=v}\left(w_{i,v}\boldsymbol{m}_{i,v}^a + \sum_{j\neq v} w_{ij}\boldsymbol{m}_{ij}^a\right) > \max_{i\in P_n} f_{j=v}\left(w_{i,v}\boldsymbol{m}_{i,v}^n + \sum_{j\neq v} w_{ij}\boldsymbol{m}_{ij}^n\right) \tag{9}$$

式中，$\boldsymbol{m}_{i,v}^a$ 代表异常视频中的示例片段，$\boldsymbol{m}_{i,j}^n$ 代表正常视频中的示例片段，$w_{i,v}$ 和 $w_{i,j}$ 分别表示示例 v 和其本身的注意力权重，以及示例 v 和包内其他示例的注意力权重，是在 LSC-transMIL 模块中习得的注意力权重。该权值代表示例 v 对应的异常分数不仅取决于其本身特征，还受相关示例片段的影响，从而降低了示例信息的不确定性，提高异常检测的准确率。

2.4　二阶段伪标签监督的时空注意力异常检测网络

为了提高视频的异常检测性能，本文将第 1 阶段生成的伪标签用于指导第 2 阶段的异常检测。该阶段的网络如图 3 所示，主要包含 3 个部分：1）Slow-Fast 骨干网络提取视频特征（Feichtenhofer 等，2019）；2）注意力模块对 Slow 和 Fast 支路视频特征进行时空注意力加权和信息融合；3）异常检测模块使用伪标签 $L_{P,i}$ 引导分类网络的训练以生成每个视频片段异常分数。网络输入为视频帧 $\hat{\boldsymbol{X}} \in \mathbf{R}^{B\times C\times T\times H\times W}$，输出为异常分数 $\hat{\boldsymbol{S}} \in \mathbf{R}^{B\times T\times 1}$。特征提取网络使用 SlowFast 网络，可以保证较高准确率的同时简化训练过程，避免提取光流这一耗时耗力的操作。

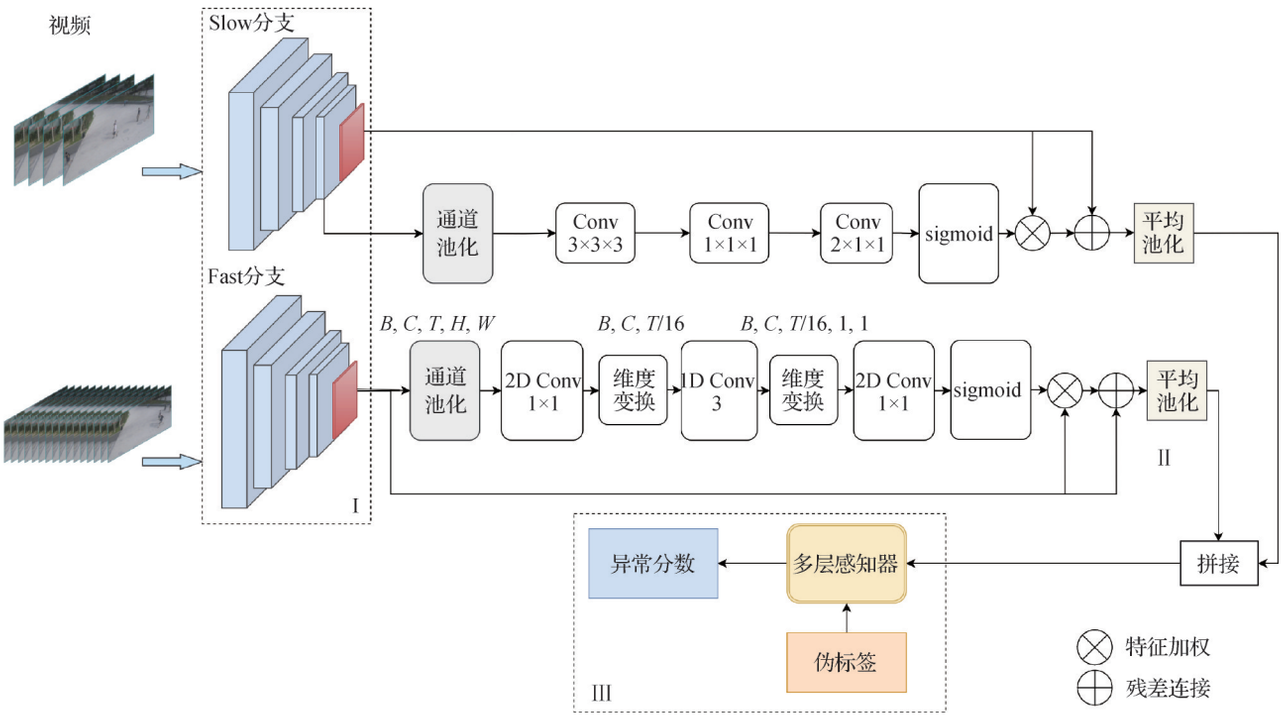


图3　伪标签监督下异常检测网络结构

Fig. 3　Anomaly detection network structure with pseudo-label

Slow支路获取视频的时空信息，通过时空注意力模块关注视频帧时空信息。来自Slow支路的输入为$\boldsymbol{X}_s \in \mathbf{R}^{B\times C\times T/8\times H\times W}$，$B$代表Batchsize，$T$代表视频帧数，$C$代表通道数，$H$，$W$代表图片高和宽。首先对所有通道进行平均，获取全局时空张量$\boldsymbol{F} \in \mathbf{R}^{B\times 1\times T/8\times H\times W}$，通过3D卷积层、sigmoid激活层获取权重矩阵$\boldsymbol{W}_s \in \mathbf{R}^{B\times 1\times T/8\times H\times W}$，$\odot$为矩阵乘法，Slow支路经时空注意力模块的输出$\boldsymbol{F}_s$为

$$\boldsymbol{F}_s = \boldsymbol{X}_s + \boldsymbol{X}_s \odot \boldsymbol{W}_s \tag{10}$$

Fast支路主要获取视频帧之间的时间相关信息。首先对输入$\boldsymbol{X}_f \in \mathbf{R}^{B\times C\times T\times H\times W}$进行空间平均池化来获取输入特征的全局信息$\boldsymbol{F} \in \mathbf{R}^{B\times C\times T\times 1\times 1}$。使用升维降维（squeeze-unsqueeze）策略（Wang等，2021）对时间维度T进行处理，使用1×1卷积将时间维度压缩为原本的1/8，经过维度变换（reshape）操作获得$\boldsymbol{F}^* \in \mathbf{R}^{B\times C\times T/8\times 1\times 1}$，经过三维卷积层、reshape、unsqueeze和sigmoid激活层获得时间权重矩阵$\boldsymbol{W}_f \in \mathbf{R}^{B\times C\times T\times 1\times 1}$，Fast支路经注意力模块输出$\boldsymbol{F}_f$为

$$\boldsymbol{F}_f = \boldsymbol{X}_f + \boldsymbol{X}_f \odot \boldsymbol{W}_f \tag{11}$$

将两条支路视频特征进行池化、展平、拼接操作后得到最终的视频特征，然后输入异常检测网络。这里将异常检测问题视为伪标签监督下的视频二分类问题，构建基于MLP的分类网络，使用sigmoid层激活，获取每一视频片段的异常分数。损失函数为交叉熵损失函数，示例标签使用第1阶段生成的细粒度标签$\boldsymbol{L}_{P,i}$，具体为

$$\hat{\boldsymbol{S}}_i = MLP_{\text{sig}}(\boldsymbol{F}_s, \boldsymbol{F}_f, \boldsymbol{L}_{P,i}) \tag{12}$$

同时，训练过程会微调SlowFast网络最后一层，减少行为识别训练集预训练模型的偏见，增强对异常事件的敏感性。

3　实验与结果

为验证所提方法的有效性，本文选用公开数据集ShanghaiTech和UCF-Crime进行实验。ShanghaiTech数据集包含13个场景中的130个异常，包含不同的拍摄角度和光照条件，其中训练视频238个，测试视频199个（Zhong等，2019）。UCF-Crime数据集是一个大型的真实监控视频数据集，包括13种异常事件，1 900个未经修剪的长视频，其中训练视频1 610个。模型训练时采用分步方式：第1阶段首先使用在Kinetics-400数据集上预训练好的参数固定的SlowFast骨干网络提取视频特征，然后基于LSC-transMIL模型生成示例伪标签；第2阶段采用伪标签监督策略训练异常检测网络，训练过程中对SlowFast骨干网络参数进行微调，使其具有更好的自适

应性。

为验证所提方法的有效性，首先对LSC-transMIL模型、全局和局部注意力模块分别进行消融实验，并对全局和局部模块的融合系数进行实验分析。然后与现有主要的弱监督异常检测方法进行比较分析，从而进一步验证所提方法的性能。依据Liu等人(2018)、Liu和Ma(2019)以及Wan等人(2020)之前的工作，将视频帧级的受试者工作特征曲线(receiver operating characteristic curve，ROC)的曲线下面积AUC作为主要度量，AUC越大意味着异常检测能力越强。

3.1 消融实验

3.1.1 LSC-transMIL模块消融实验

目前常用的提取视频帧特征的骨干网络包括I3D(inflated 3D convnet)，I3D-RGB，C3D和SlowFast。由于I3D网络光流支路计算光流需要占用更多的存储空间和时间，算法复杂度高，因此使用I3D-RGB，C3D，SlowFast 3种网络作为骨干网络对本文提出的LSC-transMIL模块进行消融实验。3种骨干网络均采用Kinetics-400数据集上的预训练模型来提取视频特征。使用C3D网络提取视频特征时，每16帧作为一个视频片段，将网络的fc6层输出作为C3D特征，特征维度为4 096。使用I3D_RGB骨干网络提取视频特征时将视频帧的最小尺寸重新缩放为256，并执行尺寸为224×224像素的中心裁剪(Zach等，2007)，每16帧作为一个视频片段生成视频特征，I3D-RGB特征维度为1 024(Carreira和Zisserman，2017)。SlowFast网络提取特征时使用官方SlowFast_r50_8×8配置文件，每连续32帧作为一个视频片段并对其进行复制，共64帧作为Fast分支输入，对这64帧进行间隔为8的采样，获得8帧图像作为Slow分支输入，最终生成的SlowFast视频特征维度为2 304。LSC-transMIL网络每批随机选取40个视频进行训练，20个正常视频和20个异常视频，使用Adagrad优化器以0.01的学习率训练生成器，设置局部全局超参量$\alpha=10$，异常阈值$\psi=0.5$。

消融实验基于这3种骨干网络测试了图1中所示的基础MIL模型(图1(a))、基于注意力机制的MIL模型(图1(b))以及本文提出的LSC-transMIL模型在ShanghaiTech数据集和UCF-Crime数据集上的效果，实验结果如表1所示。由于加强了示例间的局部和全局关联，在骨干网络相同的情况下，本文LSC-transMIL方法在第1阶段时就能获得均高于其他两种异常检测方法的AUC值，在ShanghaiTech数据集上能提升3.43%以上，在UCF-Crime数据集上能提高1%～2%。

图4显示了以SlowFast为骨干网络时3种MIL方法的ROC曲线(SlowFast特征)，在低误报率的情况下，LSC-transMIL(红色)比其他方法获得了更高的真阳性率，可以证明本文提出的LSC-transMIL方法的有效性。据表1在两个数据集上的测试表明，其他因素一致情况下，第1阶段异常检测过程中SlowFast网络提取的特征较其他骨干网络性能均为最优，即其生成的伪标签质量更高。综上，本文选择SlowFast网络作为第2阶段的特征提取网络并为其设计相应的时空注意力模块。

表1 LSC-transMIL消融实验

Table 1 LSC-transMIL ablation experiment

/%

方法	视频特征	AUC	
		ShanghaiTech	UCF-Crime
Basic MIL (Sultani等，2018)	C3D	90.00	74.33
	I3D-RGB	88.97	74.86
	SlowFast	90.12	76.17
Attention MIL (Zhu和Newsam，2019)	C3D	85.79	74.39
	I3D-RGB	89.65	75.82
	SlowFast	92.11	77.08
LSC-transMIL	C3D	94.22	79.24
	I3D-RGB	92.13	80.47
	SlowFast	**95.54**	**81.41**

注：加粗字体表示各列最优结果。

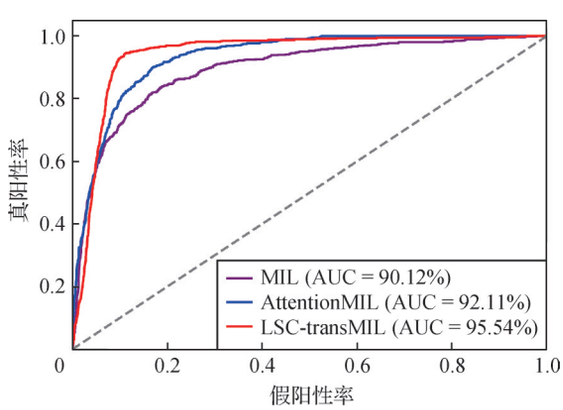


图4 3种MIL方法ROC曲线

Fig. 4 ROC curves of three MIL methods

3.1.2　全局和局部注意力消融实验

本文在第1阶段中使用全局和局部注意力机制加强示例之间的联系。为了验证其对异常检测的有效性，本文研究了这两个注意力在网络中的作用，在ShanghaiTech数据集和UCF-Crime数据集上进行实验，骨干网络为SlowFast，实验结果如表2所示。可以看出，添加局部注意力机制后可有效改善异常检测准确率，在ShanghaiTech数据集上AUC值提高了1.35%，在UCF-Crime数据集上提高了1.31%。

表2　全局和局部注意力消融实验

Table 2　Global and local attention ablation experiments

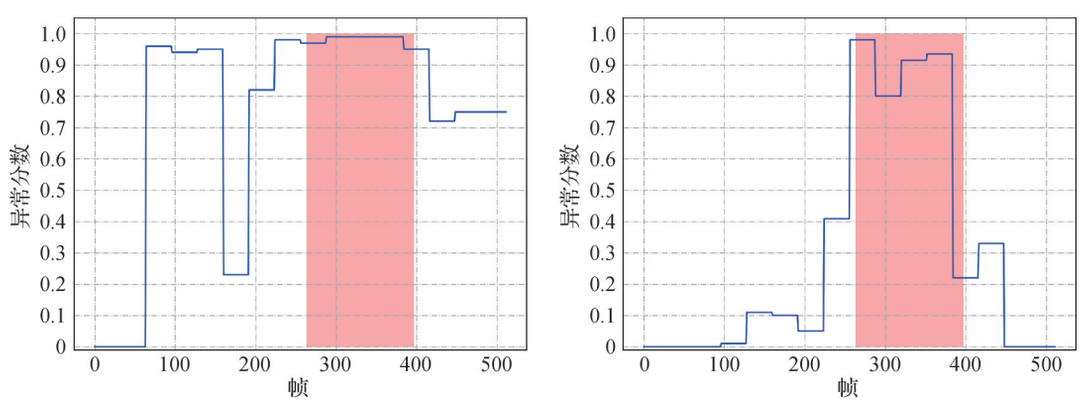
骨干网络	注意力机制	AUC/%	
		ShanghaiTech	UCF-Crimes
SlowFast	全局	94.19	80.10
	全局 + 局部	**95.54**	**81.41**

注：加粗字体表示各列最优结果。

图5和图6显示了仅有全局注意力和添加局部注意力之后，在ShanghaiTech数据集上的定性效果。其中01_0056号视频为异常视频，01_021号视频为正常视频，红色区域为异常发生的片段，蓝色线条代表模型预测的异常发生的异常分数。

如图5所示，增加局部注意力可有效改善异常检测的准确率，尤其有助于在异常视频中区分异常片段和正常片段。在仅使用全局注意力机制时，正常视频示例的异常分数均接近0，而异常视频的检测效果较差，无法区分异常视频中的正常片段和异常片段，会将正常片段误判为异常。添加局部注意力后，这种误判明显减少。

正常视频如图6所示，在局部 + 全局注意力下，强化了局部特征，使得一些特征与正常特征出现差异，略微偏离了正常特征聚集范围，异常分数出现极小范围的波动，虽然没有仅使用全局注意力时的效

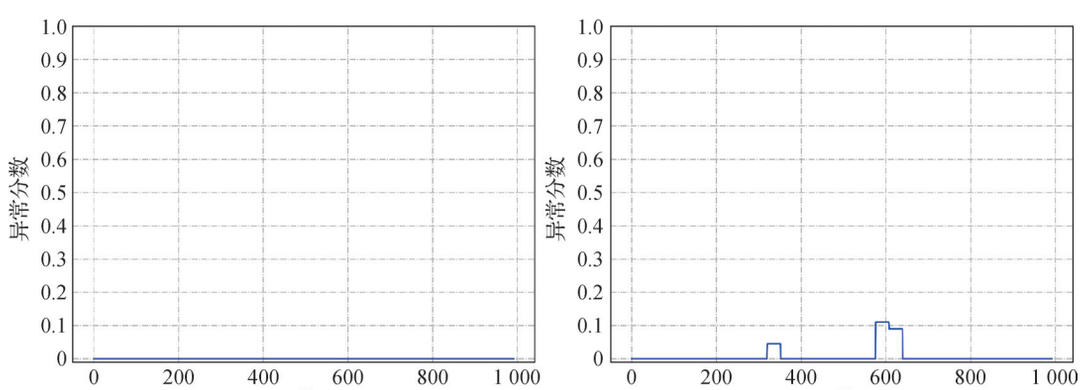


(a) 全局注意力　　(b) 局部+全局注意力

图5　ShanghaiTech数据集01_0056号视频异常检测结果

Fig. 5　Anomaly detection results of video 01_0056 of ShanghaiTech dataset

((a) global correlation attention; (b) global and local correlation attention)



(a) 全局注意力　　(b) 局部+全局注意力

图6　ShanghaiTech数据集01_021号视频异常检测结果

Fig. 6　Anomaly detection results of video 01_021 of ShanghaiTech dataset

((a) global correlation attention; (b) global and local correlation attention)

果好，但并不会影响模型的正确判断。

图 7 为 ShanghaiTech 数据集 01_0056 号视频的注意力矩阵 $\boldsymbol{A}$（式(7)）热力图，横纵坐标代表视频片段，颜色越深代表注意力矩阵的值越大。可以看出，仅使用全局注意力时，示例会与视频的其他所有示例构建联系，容易带来误判；而添加局部注意力机制后，会加强示例与临近示例之间的关注，从而减少误判。

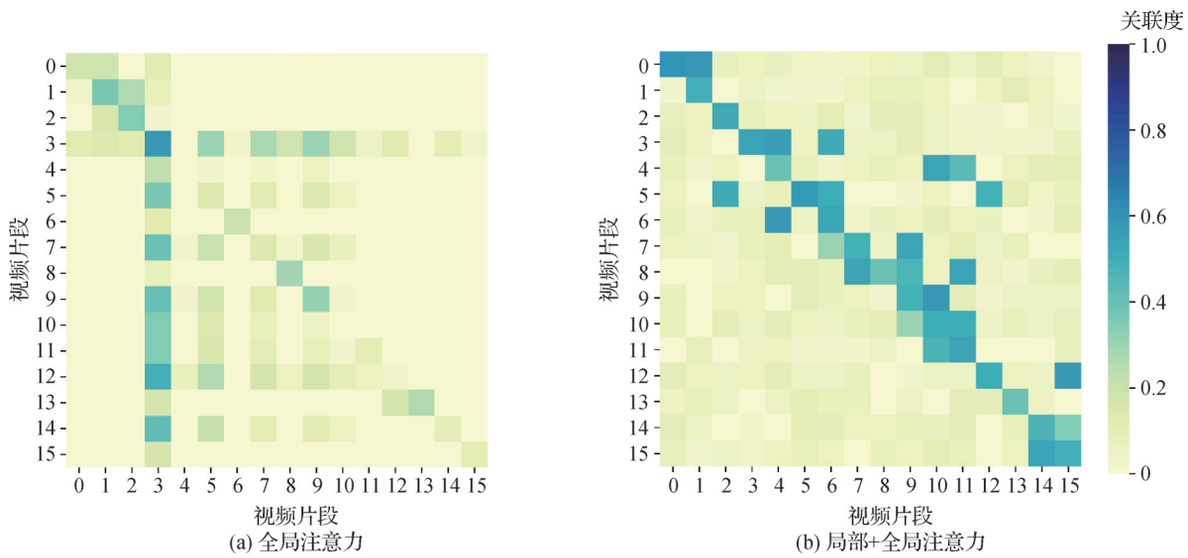


图 7　ShanghaiTech 数据集 01_0056 号视频模型注意力矩阵热图

Fig. 7　Model attention matrix heatmap of 01_0056 of ShanghaiTech dataset

((a) global correlation attention; (b) global and local correlation attention)

3.1.3　全局和局部注意力比重超参量

LSC-transMIL 模块中使用超参量 α 融合全局和局部注意力，α 取值代表对局部和全局的关注程度。通过以上消融实验可知，局部注意力可以增强网络区分正常片段与异常片段的能力，但 α 取值过大会使得示例间的全局联系变弱，信息熵降低，示例判断不确定性增大，准确率降低。因此本文通过多次实验，发现 α 取值为 10 时会获得最优的异常检测效果，比较结果如图 8 所示。

3.2　与现有异常检测方法的比较

第 2 阶段伪标签监督下异常分类网络采用 SlowFast 网络提取视频特征，相较于 I3D 网络，可以避免计算光流这一费时费力的操作，计算复杂度也有显著降低，与现有主流方法 CAC（cluster attention contrast）、GCL（generative cooperative learning）、AAVAD（anomaly attention-based weakly supervised video anomaly detection）、CNL（collaborative normality learning）、AR-Net（anomaly regression net）、MIST（multiple instance self-training framework）、MCR（multi-scale continuity-aware refinement network）和 ST-Graph（spatio-temporal context graph）在数据集 ShanghaiTech 和 UCF-Crime 比较结果如表 3 和表 4 所示，表中 Ours-one 表示本文仅进行第 1 阶段的结果，Ours-two 为本文两阶段的结果。

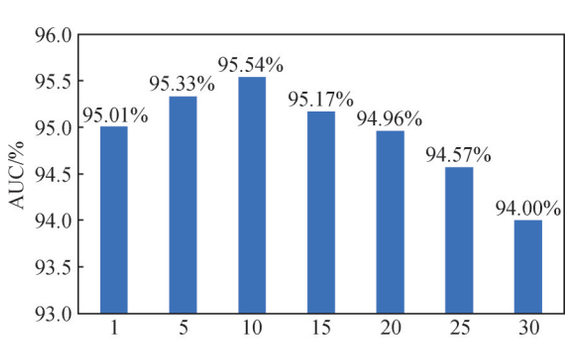


图 8　ShanghaiTech 数据集 α 不同取值对 AUC 的影响

Fig. 8　AUC of different α values on the ShanghaiTech dataset

从表 3—表 4 中可以看出，进行第 2 阶段的训练后，使用 SlowFast 视频特征提取网络在 ShanghaiTech 数据集上 AUC 达到 96.41%，在 UCF-Crime 数据集上 AUC 达到 82.88%。相较于第 1 阶段的结果分别提升了 0.87% 和 1.47%，相比于同为 2 阶段的 MIST 网络分别提升了 1.58% 和 0.58%，效果提升的一个关键因素就是 MIST 使用传统的 MIL 方法生成伪标签，而本文提出的一阶段 LSC-transMIL 方法可以生成更为可靠的细粒度伪标签。相较于多尺度连续性感知细化网络（MCR）的效果分别提高了 1.49% 和

表3 ShanghaiTech数据集上与现有方法比较

Table 3 Performance comparison on ShanghaiTech dataset

/%

方法	监督	视频特征	AUC
CAC(Wang等,2020)	Un	-	79.30
GCL(Zaheer等,2022)	Un	ResNext	78.93
Sultani等人(2018)	Weakly	C3D	86.30
Zhu和Newsam (2019)	Weakly	AE-Flow	82.50
AAVAD(Ma和Zhang,2022)	Weakly	I3D	85.70
CNL(Liu等,2022)	Weakly	-	88.20
AR-Net(Wan等,2020)	Weakly	C3D	85.01
AR-Net(Wan等,2020)	Weakly	I3D	91.24
MIST(Feng等,2021)	Weakly	C3D	93.13
MIST(Feng等,2021)	Weakly	I3D-RGB	94.83
MCR(Gong等,2022)	Weakly	I3D	94.92
Ours-one	Weakly	SlowFast	<u>95.54</u>
Ours-two	Weakly	SlowFast	**96.41**

注:加粗字体和下划线字体分别表示最优、次优的异常检测结果,"-"代表未提供此项数据,Un和Weakly分别表示无监督和弱监督。

表4 UCF-Crime数据集上与现有方法比较

Table 4 Performance comparison on UCF-Crime dataset

/%

方法	监督	视频特征	AUC
ST-Graph(Sun等,2020)	Un	-	72.70
GCL(Zaheer等,2022)	Un	ResNext	71.04
Zhang等人(2019)	Weakly	C3D	78.70
Zhu和Newsam (2019)	Weakly	AE-Flow	79.00
MCR(Gong等,2022)	Weakly	I3D	81.00
Zhong等人(2019)	Weakly	C3D	81.08
MIST(Feng等,2021)	Weakly	I3D-RGB	82.30
AAVAD(Ma和Zhang,2022)	Weakly	I3D	<u>82.60</u>
Ours-one	Weakly	SlowFast	81.41
Ours-two	Weakly	SlowFast	**82.88**

注:加粗字体和下划线字体分别表示最优、次优的异常检测结果,"-"代表未提供此项数据,Un和Weakly分别表示无监督和弱监督。

1.6%,这得益于Transformer网络强大的长序列建模能力和深层信息挖掘能力。

二阶段视频异常检测的定性结果如图9所示,对于异常帧,本文所提模型会生成高异常分数,实现较为及时的检测。对于没有异常发生的正常帧,本

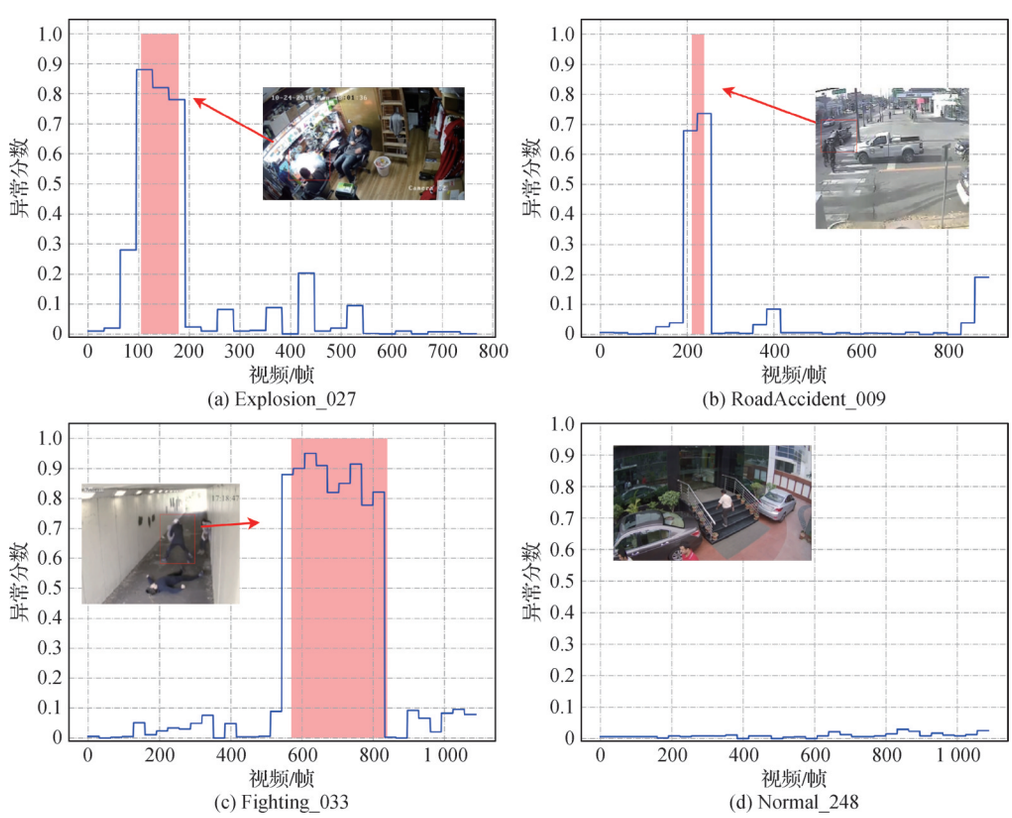


图9 UCF-Crime数据集异常检测可视化

Fig. 9 Visual examples of prediction results on UCF-Crime

((a) Explosion_027;(b) RoadAccident_009;(c) Fighting_033;(d) Normal_248)

文模型始终生成较低的异常分数。异常检测结果相较于真实的异常事件标签会存在少许超前或滞后的现象，这是由于测试集是逐帧标注的，而模型选取32帧视频帧作为一个片段生成一个异常分数。当这32帧同时包含正常和异常视频帧时，若异常在该片段后几帧开始出现，则会有以下两种情况：1）该片段判定为异常，异常检测结果相较于真实异常会少许超前，如图9（b）所示；2）该片段判定为正常，而后续片段为异常，则较于真实异常会滞后。

4 结 论

本文提出了基于MIL的LSC-transMIL-stageT方法，第1阶段LSC-transMIL模块在多示例学习过程中使用Transformer模型并加入局部和全局时间注意力加强示例之间时间联系，改善视频异常检测效果，并进行消融实验验证了LSC-transMIL模块的有效性。第2阶段使用伪标签的训练策略，使用第1阶段生成的高可信度的细粒度标签，构建由特征提取、时空注意力、视频分类3部分组成的异常检测网络，进一步提高视频异常检测准确度。

实验发现添加Transformer结构和时间维度注意力可有效提高异常检测准确率，伪标签策略有效的一大原因就是在训练过程中可以微调在动作识别数据集上预训练的模型，能够增强模型对异常事件的敏感性，但现阶段工作很难将这两种方法的优点结合起来，模型会过于复杂，接下来一个有探索意义的方向就是使用新的策略尝试将这两种方法结合起来，这将会成为下一阶段的研究重点。

参考文献（References）

Abbas Z K and Al-Ani A A. 2022. A comprehensive review for video anomaly detection on videos//Proceedings of 2022 International Conference on Computer Science and Software Engineering (CSASE). Duhok, Iraq: IEEE: #9759598 [DOI: 10.1109/CSASE51777.2022.9759598]

Carbonneau M A, Cheplygina V, Granger E and Gagnon G. 2018. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognition, 77: 329-353 [DOI: 10.1016/j.patcog.2017.10.009]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Feichtenhofer C, Fan H Q, Malik J and He K M. 2019. SlowFast networks for video recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6201-6210 [DOI: 10.1109/ICCV.2019.00630]

Feng J C, Hong F T and Zheng W S. 2021. MIST: multiple instance self-training framework for video anomaly detection//Proceedings of 2021 IEEE/CVF conference on computer vision and pattern recognition. Nashville, USA: IEEE: 14004-14013 [DOI: 10.1109/CVPR46437.2021.01379]

Gong Y L, Wang C, Dai X M, Yu S H, Xiang L H and Wu J F. 2022. Multi-scale continuity-aware refinement network for weakly supervised video anomaly detection//Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei, China: IEEE: 1-6 [DOI: 10.1109/ICME52920.2022.9860012]

Ilse M, Tomczak J M and Welling M. 2018. Attention-based deep multiple instance learning//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: JMLR: 2127-2136

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

Liang J F, Li T, Yang J Q, Li Y N, Fang Z W and Yang F. 2023. Video anomaly detection by fusing self-attention and autoencoder. Journal of Image and Graphics, 28(4): 1029-1040 (梁家菲, 李婷, 杨佳琪, 李亚楠, 方智文, 杨丰. 2023. 融合自注意力和自编码器的视频异常检测. 中国图象图形学报, 28(4): 1029-1040) [DOI: 10.11834/jig.211147]

Liu K and Ma H D. 2019. Exploring background-bias for anomaly detection in surveillance videos//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1490-1499 [DOI: 10.1145/3343031.3350998]

Liu W, Luo W X, Lian D Z and Gao S H. 2018. Future frame prediction for anomaly detection——a new baseline//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6536-6545 [DOI: 10.1109/CVPR.2018.00684]

Liu Y, Liu J, Zhao M Y, Li S and Song L. 2022. Collaborative normality learning framework for weakly supervised video anomaly detection. IEEE Transactions on Circuits and Systems II: Express Briefs, 69(5): 2508-2512 [DOI: 10.1109/TCSII.2022.3161061]

Ma H L and Zhang L Y. 2022. Attention-based framework for weakly supervised video anomaly detection. The Journal of Supercomputing, 78(6): 8409-8429 [DOI: 10.1007/s11227-021-04190-9]

Majhi S, Dash R and Sa P K. 2020. Temporal pooling in inflated 3DCNN for weakly-supervised video anomaly detection//Proceed-

ings of the 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Kharagpur, India: IEEE: 1-6 [DOI: 10.1109/ICCCNT49239.2020.9225378]

Shao Z C, Bian H, Chen Y, Wang Y F, Zhang J, Ji X Y and Zhang Y B. 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification [EB/OL]. [2023-06-28]. https://arxiv.org/pdf/2106.00908.pdf

Shi X S, Xing F Y, Xie Y P, Zhang Z Z, Cui L and Yang L. 2020. Loss-based attention for deep multiple instance learning//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 5742-5749 [DOI: 10.1609/aaai.v34i04.6030]

Sultani W, Chen C and Shah M. 2018. Real-world anomaly detection in surveillance videos//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6479-6488 [DOI: 10.1109/CVPR.2018.00678]

Sun C, Jia Y D, Hu Y and Wu Y W. 2020. Scene-aware context reasoning for unsupervised abnormal event detection in videos//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 184-192 [DOI: 10.1145/3394171.3413887]

Wan B Y, Fang Y M, Xia X and Mei J J. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning// Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE: #9102722 [DOI: 10.1109/ICME46284.2020.9102722]

Wang Z G and Zhang Y J. 2020. Anomaly detection in surveillance videos: a survey. Journal of Tsinghua University (Science and Technology), 60(6): 518-529 (王志国, 章毓晋. 2020. 监控视频异常检测: 综述. 清华大学学报(自然科学版), 60(6): 518-529) [DOI: 10.16511/j.cnki.qhdxxb.2020.22.008]

Wang Z W, She Q and Smolic A. 2021. ACTION-Net: multipath excitation for action recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13209-13218 [DOI: 10.1109/CVPR46437.2021.01301]

Wang Z M, Zou Y X and Zhang Z M. 2020. Cluster attention contrast for video anomaly detection//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2463-2471 [DOI: 10.1145/3394171.3413529]

Wei X S and Zhou Z H. 2016. An empirical study on image bag generators for multi-instance learning. Machine Learning, 105(2): 155-198 [DOI: 10.1007/s10994-016-5560-1]

Zach C, Pock T and Bischof H. 2007. A duality based approach for real-time TV-L^1 optical flow//Proceedings of the 29th DAGM Symposium on Pattern Recognition. Heidelberg, Germany: Springer: 214-223 [DOI: 10.1007/978-3-540-74936-3_22]

Zaheer M Z, Mahmood A, Khan M H, Segu M, Yu F and Lee S I. 2022. Generative cooperative learning for unsupervised video anomaly detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14724-14734 [DOI: 10.1109/CVPR52688.2022.01433]

Zhang J G, Qing L and Miao J. 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 4030-4034 [DOI: 10.1109/ICIP.2019.8803657]

Zhong J X, Li N N, Kong W J, Liu S, Li T H and Li G. 2019. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1237-1246 [DOI: 10.1109/CVPR.2019.00133]

Zhou H, Zhan Y Z and Mao Q R. 2021. Video anomaly detection based on space-time fusion graph network learning. Journal of Computer Research and Development, 58(1): 48-59 (周航, 詹永照, 毛启容. 2021. 基于时空融合图网络学习的视频异常事件检测. 计算机研究与发展, 58(1): 48-59) [DOI: 10.7544/issn1000-1239202120200264]

Zhu Y and Newsam S. 2019. Motion-aware feature for improved video anomaly detection [EB/OL]. [2023-06-28]. https://arxiv.org/pdf/1907.10211.pdf

作者简介

朱新瑞,男,硕士研究生,主要研究方向为视频异常检测。
E-mail: xinruizhu@nuaa.edu.cn
钱小燕,通信作者,女,副教授,主要研究方向为深度学习和智能监控。E-mail: qianxiaoyan@nuaa.edu.cn
施俞洲,男,硕士研究生,主要研究方向为小目标检测。
E-mail: shiyuzhou@nuaa.edu.cn
陶旭东,男,硕士研究生,主要研究方向为目标跟踪。
E-mail: taoxudong0708@nuaa.edu.cn
李智昱,男,硕士研究生,主要研究方向为深度学习和图像处理。E-mail: lizhiyu@nuaa.edu.cn