文章编号:1671-4598(2025)05-0045-08 DOI:10.16526/j.cnki.11-4762/tp.2025.05.006 中图分类号:TP391 文献标识码:A

基于多尺度时空混合专家增强的视频异常检测方法

王宁冰，赵佳佳，陈国辉
（长沙理工大学 计算机与通信工程学院，长沙 410114）

摘要：针对传统视频异常检测技术中异常数据稀缺导致的对视频特征理解不足，多异常环境下适应性较差的问题，提出了基于多尺度时空混合专家增强的视频异常检测方法；具体来说，使用 I3D 提取视频特征，从空间和时间两个维度下对视频特征进行多尺度增强，提升对视频特征的理解能力；利用混合专家模块对增强的多尺度特征进行异常得分回归，增强多类型异常检测的鲁棒性；同时，利用多尺度时空对应关系建立回归损失函数，进一步提高了检测精度；在 UCF-Crime 数据集和 ShanghaiTech 数据集上的实验结果表明，所提出方法的视频的异常检测准确率有明显提升，多异常场景下检测效果较为稳定，满足实际需求。

关键词：视频异常检测；弱监督学习；多尺度时空特征信息增强；混合专家模型

Enhanced Video Anomaly Detection Method Based on Multi-scale Spatio-temporal Hybrid Experts

WANG Ningbing, ZHAO Jiajia, CHEN Guohui
(School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China)

Abstract: Aiming at the problems of insufficient understanding of video features and poor adaptability in multi-anomaly environments caused by scarce abnormal data in traditional video anomaly detection techniques, an enhancement video anomaly detection method based on multi-scale spatio-temporal hybrid experts is proposed; Specifically, using inflated 3D convnet (I3D) to extract video features, perform multi-scale enhancement of video features from spatial and temporal dimensions, and improve the comprehension of video features; a hybrid expert module is used to perform anomaly score regression on the enhanced multi-scale features, and enhance the robustness of multi-type anomaly detection; At the same time, a regression loss function is established using multi-scale spatio-temporal correspondence, which further improves detection accuracy; Experimental results on the UCF-Crime dataset and ShanghaiTech dataset show that the proposed method significantly improves the accuracy of video anomaly detection, and the detection effect is relatively stable in multiple anomaly scenarios, which meets practical requirements.

Keywords: video anomaly detection; weak supervised learning; enhancement of multi-scale spatio-temporal feature information; hybrid expert model

0 引言

随着监控摄像头广泛应用，处理大量视频监控数据已使人工检测无法满足需求。因此，基于人工智能的视频异常检测变得至关重要[1-3]。在视频异常检测中，异常通常定义为与正常行为或场景相比较显著不同的行为或场景[4]。近年来，弱监督视频异常检测方法因只需对训练数据进行视频级别标注，并且仅使用包含异常事件的少量视频即可完成训练的特点，降低了数据标注成本，同时其训练效果往往优于无监督学习，成为了当前研究的热点[5-6]。

当前弱监督视频异常检测主要依赖文献［7］提出的多示例学习（MIL，multiple-instance learning）框架

收稿日期:2024-03-02； **修回日期**:2024-04-24。
基金项目:国家自然科学基金项目(62371278,62371279)。
作者简介:王宁冰(1998-),男,硕士研究生。
通讯作者:赵佳佳(1988-),女，博士，讲师，CCF 会员。
引用格式:王宁冰,赵佳佳,陈国辉. 基于多尺度时空混合专家增强的视频异常检测方法[J]. 计算机测量与控制,2025,33(5):45-52.

的异常检测算法。在该框架下，视频异常检测效果获得了显著提升，然而，由于异常动作通常是连续且稀少的，现有方法容易受到视频中正常片段的噪声干扰。同时，多示例学习框架针对正包和负包中的片段进行打分，使得模型最终得分主要受到某几个异常片段的影响，这使得模型学习速度慢、泛化困难。此外，单一模型在面对复杂异常场景时鲁棒性受限，再加上受限于样本数量的稀缺性，已有模型的识别效果与理想准确率之间存在一定的差距。这些问题表明当前弱监督视频异常检测方法在处理异常视频时仍有改进的空间[8-10]。在视频处理和分析的领域，时间卷积已被广泛认可为一种高效的特征增强技术。文献［11］基于时间卷积的原理，进一步提出了一种创新的优化策略，即通过引入相邻帧的信息来改善特征表示。这种策略在提高特征质量方面表现出了显著的效果。文献［12］将研究重点放在了噪声数据的处理上。他们采用了图卷积网络来纠正噪声标签，从而确保了数据在时空维度上的一致性。文献［13］开始对 MIL 的损失函数优化，近期的研究工作开始聚焦于多实例学习框架的优化[14]。这些研究努力在提升框架性能和效率方面取得了一定的成果。然而这些方法在处理异常数据稀缺问题时，并未给予足够的关注，同时对于视频特征理解也未达到应有的深度。

为了应对上述问题，本文提出了一种多尺度时空混合专家增强的视频异常检测方法。通过整合多尺度的时空信息，引入混合专家机制，该方法旨在提高模型对视频时间空间语义的理解，同时解决复杂异常场景下模型鲁棒性的问题。

1 视频异常检测模型

1.1 整体框架

本文提出的多尺度时空混合专家增强弱监督视频异常检测方法如图 1 所示，包含了特征提取、多尺度时空信息增强、混合专家增强 3 个部分。首先，通过使用一个经过训练的 I3D 网络从视频中提取特征。接着，利用多尺度时空特征增强模块对视频特征进行增强和划分。随后，在此基础上，将视频异常得分网络定义专家模型，采用（MoE，mixture of experts）[15]架构，利用门控网络对多个专家模型的性能和置信度进行动态调整，最终得到该视频的异常得分。在训练时，借助多尺度卷积模块学习重建的视频特征，并使用多个损失函数对模型进行监督。

1.2 时空增强模块

时空增强模块主要包括 3 个部分，分别是多尺度时间特征模块、多尺度空间特征模块和多尺度特征融合得分模块。为了利用时空增强模块对特征进行增强，首先，输入一个标记好正常或异常的含有 T 帧视频 V，将 V 输入到预训练好的特征提取网络（I3D，inflated 3D convnet）[16]中。每 16 帧得到一个特征得分 x_i（1 * 1 024），共 m 个，其中 m 设置为 32 对于含有异常片段的视频，将其作为多示例学习中的正包，表示为 $\boxed{Vp=(P1, P2, \cdots, P32)}$；不含有异常片段的视频作为负包，表示为 $\boxed{Vn=(N1, N2, \cdots, N32)}$，如图 2 所示。

1）多尺度时间特征模块：

视频特征序列 $\boxed{Vp}$ 和 $\boxed{Vn}$ 只能在单一的时间尺度上体现视频语义。然而，视频中的异常往往是随机出现的，且具有不同的时间长度。对于一个完整的异常事件而言，仅考虑其中的某一部分可能会影响判断结果。以抢劫事件为例，包含了冲向被抢者和拿走被抢者物品两个阶段，若只考虑其中某一部分，则无法完整识别异常。然而，考虑过长的片段可能会导致片段中的正常部

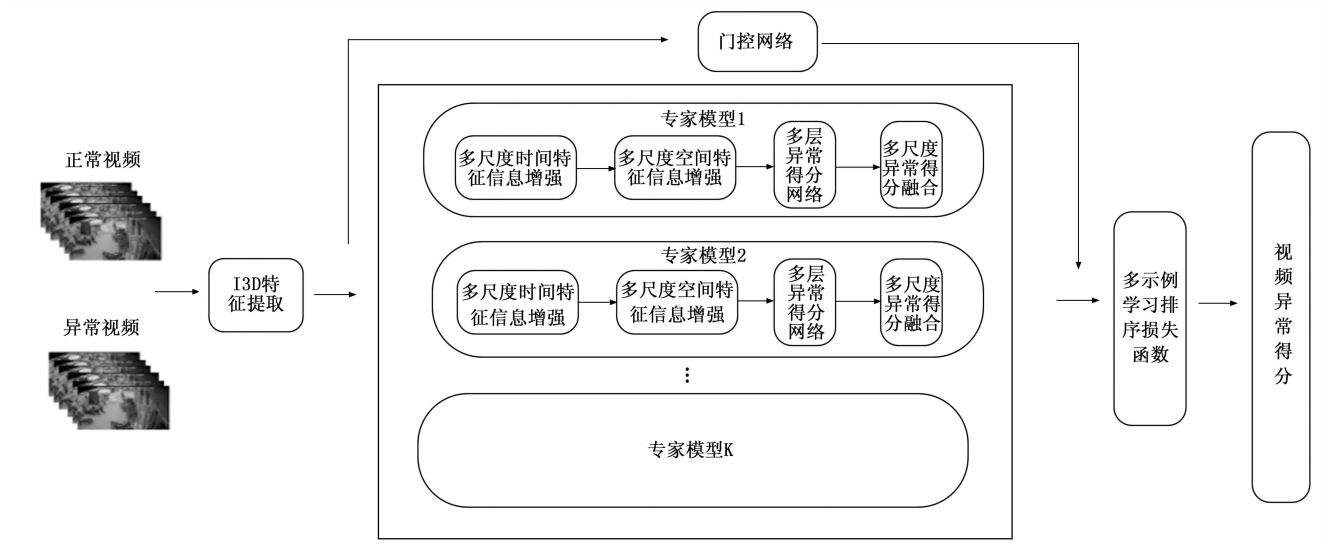

图 1 多尺度时空信息增强模块的架构

分对视频异常本身造成较大噪声，从而降低视频的准确率。

因此，我们设计了一个多尺度时间特征融合模块来生成具有多个时间尺度的特征。通过对视频特征序列 $\boxed{Vp}$ 和 $\boxed{Vn}$ 中的每一个序列进行两次平均池化，计算公式如式 1 所示。其中，每一个序列均得到了 3 个不同尺度的视频特征序列 $V_{s1}=[x_1, x_2\cdots\cdots\cdot x_{32}]^T$，$V_{s2}=[x_1, x_2\cdots\cdots\cdot x_{16}]^T$，$V_{s3}=[x_1, x_2\cdots\cdots\cdot x_8]^T$：

$$X^i=\begin{cases}x_V, & \text{if } i=1\\ \text{AveragePool}(x^{i-1}), & \text{else}\end{cases} \tag{1}$$

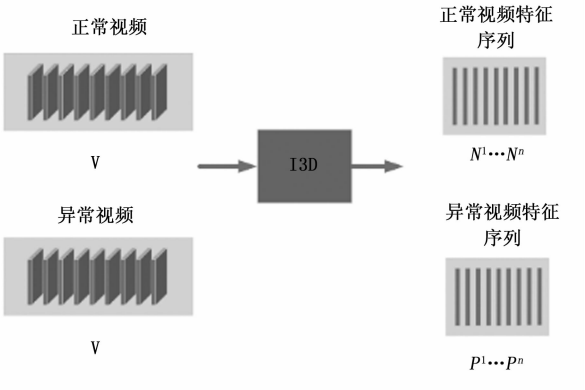

图 2　视频特征预处理的结构

3 个不同长度的时间尺度代表了从不同角度观察视频中的信息，如图 3 所示。较小的尺度可以捕捉到视频中短暂的异常情况，而较大的尺度可以确认异常情况是否已经完全发生或只是有发生的趋势。通过多个时间尺度，模型可以更全面地捕捉视频特征在不同时间尺度下的变化，确保系统能够在不同的情况下和不同的强度水平下准确地识别异常情况，提高对异常事件的敏感性和准确性。这种多尺度设计有助于提高检测的效果，增强系统的整体可靠性。

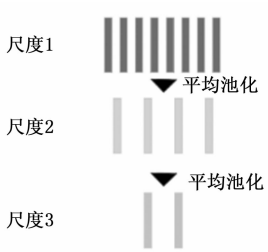

图 3　多尺度时间特征模块

2）多尺度空间特征模块：

从空间的角度看，视频特征的输入类似于人类观看视频的方向。在同一时间点，观看视频的方向会影响人类对事件的判断。同一个视频更改感兴趣区域（ROI, region of interest）[20] 会引导观察者关注不同区域，从而得到不同的观察结果。例如，一个视频中的异常发生在画面的左上方，而视频的 ROI 被定义在右下方，左上方异常波动对视频异常判断的影响将变得微小，因此多次不同的 ROI 区域划分可以有效提高模型对视频的理解能力。

在基于深度学习的视频异常检测中，提取的视频特征也包含了对应的 ROI 信息，视频特征序列对于当前视频空间内容只有一个维度的表示。为了提高特征序列的语义丰富性，从而提升模型对训练视频的理解能力，本论文设计了一个多尺度空间特征融合模块，用于增加每个尺度下特征对于空间特征的表达，如图 4 所示。具体操作是对 3 个尺度下的视频片段特征使用两个不同的全连接层进行空间卷积，这样的操作使得每个尺度下的特征都能更充分地表达空间信息，从而得到 $V_{sp1}=[x_1, x_2\cdots\cdots\cdot x_{32}]^T$，$V_{sp2}=[x_1, x_2\cdots\cdots\cdot x_{16}]^T$，$V_{sp3}=[x_1, x_2\cdots\cdots\cdot x_8]^T$。

视频中的 3 个尺度代表了视频特征的不同空间维度。通过对视频的空间维度进行多次卷积学习，能够更好地捕捉到画面中的信息变化，提高了对于不同方向的视频异常的检测能力。

在视频异常检测中使用多种时空尺度特征极大地促进了其有效性和效率。这种设计增强了系统识别异常的能力，并对视频的内容有了更全面的了解。多尺度特征的引入使得模型能够更全面地考虑空间维度上的变化，从而提高了系统对异常事件的敏感性和准确性，为视频异常检测提供了更强的表达能力。

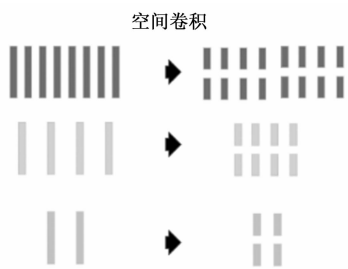

图 4　多尺度空间特征模块

3）多尺度特征融合得分模块：

异常行为由于其主观性很难准确定义，可能并不明显。本文将异常的识别视为回归问题，期望异常片段具有比正常片段更高的分数，即认为异常片段的得分应当比正常片段更高：

$$f(V_p)>f(V_n) \tag{2}$$

其中：V_p 和 V_n 表示异常和正常视频片段，$f(V_p)$ 和 $f(V_n)$ 分别表示从 0 到 1 的相应预测异常分数。

在多示例学习中，将含有异常片段的视频作为正包，$V_p=(P^1, P^2\cdots P^{32})$，不含有异常片段的视频作为负包，$V_n=(N^1, N^2\cdots N^{32})$ 输入时，正包中的得分

最高的片段应高于负包中的片段：

$$\max_{i\in V_p} f(V_p^i) > \max_{i\in V_n} f(V_n^i) \tag{3}$$

通过一个三层卷积层对 3 个尺度下的得分进行卷积，得到每个片段的得分。在 3 个尺度下选择各个部分的最大值，并进行融合，最终得到该视频的异常得分。通过网络训练约束函数对模型进行训练，从而得到有效的异常得分模块。

1.3 多专家模型层

当应用反向传播训练单个多层网络在不同情况下执行不同的子任务时，通常会导致学习缓慢且泛化能力差的强烈干扰效应。实际情况中，异常类型是多种多样的，而专家模型网络的优势在于可以针对不同的训练数据训练出拥有多个不同侧重点的专家模型。这实现了专家与擅长类型之间的一对多关系，最终的视频异常得分由多个专家参与计算，提高了得分的置信度。

如果事先知道一组训练示例可以自然地划分为对应于不同子任务的子集，可以通过使用由几种不同的“专家”网络组成的系统，并使用门控网络来决定每个训练案例应该使用哪些专家，从而减少干扰。在本文中，采用了 MoE（Mixture of Experts）思想，在网络中引入多专家模型层，该模型包含 K 个专家，每个专家模包含了完整的多尺度空间特征模块和异常得分网络。通过让 K 个模型参与到专家模型层中，视频片段会得到 K 个得分，从而提高了整体模型的性能和泛化能力。K 的取值依据具体数据集的情况确定，较大的 K 值即较多的专家模型能够对更多类型的异常有适应能力，但会导致计算资源的增加。如何保证增加的模型层也能发挥应有的效果成为一个挑战。在多专家模型中，认为每个专家对于不同类型的异常视频有着不同的识别能力，训练模型时门网络将会给每个专家分配的权重，这些权重最终影响得分。如果输出不正确，权重变化会通过反向传播回专家模型层和门控网络，实现模型的优化和学习。

设 $f(x)$ 为视频异常得分函数，则对于视频片段 v，第 K 个专家的视频异常得分为 S_n，如公式（4）所示：

$$S_n = f_K(v) \tag{4}$$

设 $g(x)$ 为专家机制中的门控网络，对于视频 v，当专家个数为 K 时，可通过 $g_K(x)$ 获得每个专家的得分权重。$G(x)$ 函数的主要效果是将输入的信息划分成 K 个子集合，针对输入样本 v，生成长度为 K 的权重向量，如公式（5）所示：

$$g_K(v) = \frac{e^K}{\sum_{i=1}^{K} e^K} \tag{5}$$

1.4 网络训练的约束函数

本方法中损失函数为 $L=\lambda_1 l_{sp}+\lambda_2 l_{sm}+\lambda_3 l_{ce}+\lambda_4 l_{kl}+\lambda_5 l_{cp}+\lambda_6 l_{av}$，其中 l_{sp}、l_{sm}、l_{ce}、l_{kl}、l_{cp}、l_{av} 分别代表了稀疏损失函数、光滑损失函数、交叉熵损失函数、多尺度时空拟合损失函数、对抗损失函数和均衡损失函数。本文中，损失函数系数 λ_1、λ_2、λ_3、λ_4、λ_5、λ_6 均为超参数。

1.4.1 任务相关

1）稀疏损失函数：

在视频监控中，异常片段通常只会短暂出现，因此即使在一段异常视频中，异常实例也应该是稀疏的。过于敏感的模型会导致较多的误判，从而增加假正率。为了保证异常片段在异常视频中的稀疏性，我们在模型中增加了稀疏函数来约束模型对于异常的敏感性。这样，我们就可以在一定程度上忽略视频中一些重复性的特征，例如车灯闪烁，从而提高真阳性率（TPR，true positive rate）。具体公式如下：

$$l_{sp} = \lambda_1 \sum_{t=1}^{T}\sum_{i=1}^{C} | f_{t,i} + \lambda_2 \sum_{t=1}^{T-1}\sum_{i=1}^{C} | f_{t+1,i} - f_{t,i} | \tag{6}$$

其中：λ_1 和 λ_2 分别是控制空间和时间稀疏性的系数，T 是视频的帧数，C 是特征向量的维度，$f_{t,i}$ 表示第 t 帧第 i 个特征的值。通过这样的约束函数，模型在学习过程中会更加注重于捕捉视频中显著的异常特征，提高了模型对异常的识别准确性。

2）光滑损失函数：

对于视频帧而言，异常发生时应是一个连续的动作。例如，抢劫事件应包含抢劫者从正常行走加速冲向被抢者的过程。因此，在异常发生的片段中，视频帧的异常得分应该是连续增加的。然而，由于视频本身可能存在的一些问题，例如来源相机的成像质量或光线的变化（如夜晚灯光突然开启），会导致异常得分突变，但这些突变通常并不代表真正的异常发生。因此，我们通过增加平滑函数来对视频得分进行平滑约束，以防止分数突变，使得分数更具有连续性。具体公式如下：

$$l_{sm} = \sum_{t=1}^{T} \| f_{t+1} - f_t \|^2 \tag{7}$$

其中：f 是视频的帧数，f_t 表示第 t 帧经过特征提取后得到的特征向量。通过引入平滑函数，模型在训练时会更加注重于学习连续的异常动作特征，提高了模型对真实异常的识别准确性。

3）交叉熵损失函数：

交叉熵是一种衡量预测值与真实值之间的差异的方式，公式如下：

$$l_{ce}(x) = -\sum_{i=1}^{n} \log P(y_i = l_i \mid x_i) \tag{8}$$

其中：y_i 为异常视频的标签，p_i 为模型预测的视频的标签。x_i 为第 i 个输入 y_i 为预测得分，l_i 为真实得分，n 为样本数。

1.4.2　时空增强模块相关

多尺度时空拟合损失函数：

在进行多尺度时间特征和空间特征融合后，获得了 3 个不同尺度的视频片段特征。在这 3 个尺度中，尺度 3 中的每个片段对应的特征是 3 个尺度中最大的。通过视频尺度融合，不同尺度的视频片段在本质上是对原视频片段的不同语义表述。因此，在对每个视频包进行评分时，选择了在时空特征融合后的尺度 3 中评分最高的片段。这个片段是由尺度 1 中的 4 个片段池化为尺度 2 中的两个片段，再池化到尺度 3 得出的。因此，评分最高的片段对应着两个尺度 2 片段和 4 个尺度 1 片段。对这 3 个尺度的片段分别打分并取平均值，得到了尺度 3 的片段对应的尺度 2 与尺度 1 片段的得分，分别命名为 max _ score3、max _ score2、max _ score1。然后对这 3 个评分进行 KL 散度的接近性检测，以确定它们的相对差异。公式如下：

$$l_{kl}(x_i) = \sum q(x_i)\log\frac{q(x_i)}{p_1(x_i)} + \sum q(x_i)\log\frac{q(x_i)}{p_2(x_i)} \tag{9}$$

x_i 为输入的片段，q（x_i）为分成 8 个片段的预测得分，即 max _ score3；p_1（x_i）为分成 16 个片段的预测得分，即 max _ score2；p_2（x_i）为分成 32 个片段的预测得分即 max _ score1。通过 KL 散度的计算，可以量化不同尺度的得分之间的相对差异，为模型性能的提升提供了指导。在本论文的损失函数中，使用了两次本公式，第一次 p（x）是 max _ score3 的得分，q（x）是 max _ score1 的得分，第二次 p（x）是 max _ score3 的得分，q（x）是 max _ score2 的得分。

1.4.3　多专家模型层相关

1）对抗损失函数：

由于不同专家模型之间的相互影响较大，一旦一个专家模型的参数发生变化，其他专家模型也会相应变化，导致使用了过多的专家模型来处理样本。为了解决这个问题，引入了一个新的损失函数，定义如公式（10）所示：

$$l_{cp} = -\log\sum_K p_K^v\, e^{-1/2\,\|d^v - o_K^v\|^2} \tag{10}$$

其中：o_K^v 是样本 v 由第 K 个专家模型的得分，理想的得分为 d^v，p_K^v 是门控网络分配给第 K 个专家的权重。这个损失函数旨在促使专家模型之间进行竞争，使得每个专家模型在处理样本时发挥其特长，而不是过度依赖其他专家模型。

2）均衡损失函数：

不同专家模型在竞争过程中可能出现“赢者通吃”的现象，即在前期表现较好的专家模型更容易被门控网络选择，导致最终只有极少数几个专家模型真正发挥作用。为了缓解这种不平衡现象，本文额外引入了一个均衡损失函数，确保更多的专家模型能够被充分利用，提高模型的泛化性，具体定义如公式（11）、（12）所示：

$$\text{Importance}(v) = \sum_{x\in X} g_K(v) \tag{11}$$

$$l_{av} = \lambda \cdot CV[\text{Importance}(v)]^2 \tag{12}$$

其中：对于一个样本 v，把 v 由所有的门控网络得分 g_K（v）加起来，然后计算变异系数（CV，coefficient of variation）。变异系数 CV 反映了不同 experts 之间不平衡的程度。

2　实验

在评估模型性能时，对测试集视频按照视频帧进行标记。通过使用 I3D-RGB 网络提取测试视频的特征，直接对这些特征进行异常得分计算。随后，将得到的异常得分扩充为 16 份，与标记对齐。通过对扩充后的异常得分进行计算，得出模型在测试集上的 AUC 指标。值得注意的是，在本次实验中，损失函数系数均被设定为 1，以保持各项损失函数在训练过程中的相对权重平衡。

2.1　数据集

UCF- Crime[7] 数据集是一个包含 128 小时视频的大型数据集。它由 1 900 个连续的未经剪辑的监控视频组成，平均帧率是 7 274 帧，共包含 13 种现实生活中的异常行为，分别是虐待、逮捕、纵火、攻击、道路事故、入室盗窃、爆炸、战斗、抢劫、射击、偷窃、入店行窃和破坏。这些异常行为视频之所以被选择是因为它们对公共安全有影响。

ShanghaiTech[17] 数据集是用于异常事件检测的标准数据集。训练集有 330 个视频片段包含 27 万多个训练帧，测试集有 107 个视频片段包含 4.2 万多个测试帧（异常帧约 1.7 万个），覆盖了 13 个不同的场景。异常场景包括除行人（如车辆）以外的物体和剧烈运动（如打斗和追逐）。

2.2　评价指标

本论文采用弱监督视频异常检测领域中常用的 AUC（AUC，area under curve）[19] 指标对模型帧级别异常检测能力进行评估。AUC 被即为 ROC（ROC，receiver operation characteristic）[18] 曲线下与坐标轴围成的面积，是一个［0，1］间的分数，其值越高表示模型的异常检测性能越好。计算过程中，对测试分数集合进行排序并依次递进选择一个分数作为异常分数阈值，根据阈值计算真阳性率和假阳性率画出接受者操作特性曲线。

2.3　实验结果与分析

2.3.1　结果分析

为了评估本文提出的方法在异常检测效果上的提升，我们将其与现有的基于弱监督视频异常检测的方法进行了比较。表 1 和表 2 分别列出了在不同数据集上，当 $K=5$ 时使用不同方法得到的帧级 AUC。在

UCF-Crime 数据集上，相比当前最优的 RTFM[14] 方法，我们的方法获得绝对提升 1.3%，达到了 85.0% 的 AUC，印证了本文所提的基于多尺度混和专家增强的方法是有效的，相比基于 C3D 特征提取的方法则具有更明显的优势，这也说明了本文的多尺度时空信息增强模块的重要性。具体的 ROC 曲线图如图 5 所示。而在 ShanghaiTech 数据集上，AUC 达到了 98.5%，这证明了所做的改进对于异常事件的检测效果确实具有显著的提升。

表 1 UCF-CRIME 数据集上不同视频异常检测算法性能对比

模型	特征提取方法	AUC /%
* Lu et al.[19] Sultani et al.[7]	C3D RGB	65.51
	I3D RGB	70.46
	C3DRGB	75.41
Zhang et al.[11]	I3D RGB	78.95
GCN-Anomaly[18]	C3D RGB	81.08
MIST[16]	I3D RGB	82.30
RTFM[14]	I3D RGB	84.03
Ours	I3D RGB	85.01

* 为无监督视频异常检测

表 2 ShanghaiTech 数据集上不同视频异常检测算法性能对比

模型	AUC	AUC/%
GCN Anomaly[18]	C3D-RGB	76.44
Zhang et al.[11]	I3D-RGB	82.50
Chang et al.[21]	I3D Flow	92.25
RTFM[14]	I3D-RGB	97.21
Ours	I3D-RGB	98.51

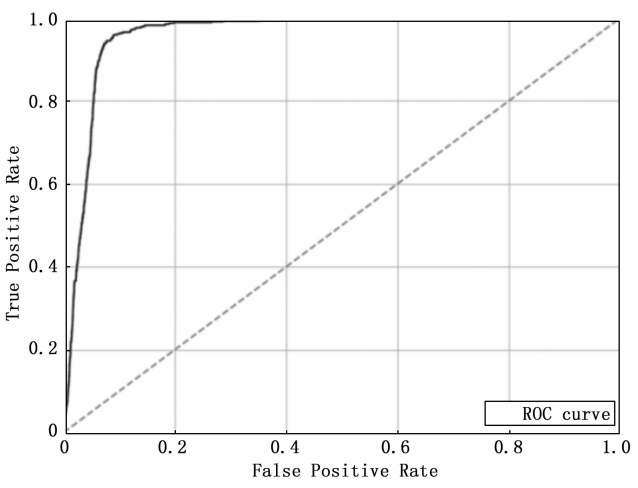

图 5 在 UCF-Crime 数据集上的 ROC 曲线

值得注意的是，我们的方法在 UCF-Crime 数据集和 ShanghaiTech 数据集上均优于先前的无监督学习方法，并且在与当前主要的基于多示例学习的弱监督异常检测方法的比较中也取得了更好的效果。这表明我们的方法在处理视频异常检测问题上具有较高的性能和有效性。

通过融合多尺度时空特征信息的方法，我们在一定程度上解决了图像噪声和光照变化等问题，从而提高了模型的鲁棒性。这种方法充分利用了视频的多个特征，进而提高了异常检测的准确性和精度，相较于基线方法在检测准确性上有了显著的提升。结果进一步表明，基于混合专家机制的模型在泛化能力上优于融合多尺度时空信息增强的模型，证明了混合专家机制的有效性。

为了说明时间融合模块和空间融合模块对模型效果的提升，我们对 UCF-Crime 数据集进行了消融实验。结果表明，时间信息增强模块和空间信息增强模块对视频效果的提升都起到了积极的作用，其中时间信息的增强对于基于 I3D 提取的特征表现出了更好的增强效果。这强化了我们方法中各模块的有效性，特别是时间信息增强模块对于视频特征的改进在提升模型性能方面具有重要作用。

表 3 模型中各个模块对模型性能的提升对比

模型	AUC/%
基线方法	77.92
基线方法＋时间信息增强	82.26
基线方法＋空间信息	78.89
基线方法＋时空信息增强	84.02
基线方法＋时空信息增强＋多专家模型	85.01

通过对异常检测模型进行改进，我们观察到在增加时间信息和空间信息的情况下，相较于基线方法，异常检测模型的性能提升了 4.34% 和 0.97%。其中，时间信息的增强对使用 I3D 提取的视频特征的提升效果更为显著，为模型性能的提升做出了更大的贡献。这可以归因于基线方法主要基于单一视频帧，未充分关注时间信息，而时间信息在异常判断中起着不可或缺的作用。通过增加时间信息，模型能够更全面地判断异常情况，类似于过去方法中对运动流信息学习的关注。而我们的方法使用了 I3D 提取的视频信息，结合时间信息增强，使得操作更为便捷，同时能够与视频提取过程中的时间信息增强相结合，共同提高视频异常检测的准确率。在引入了多专家模型的情况下，异常识别的准确率提升了 0.99%。多专家的差异化使得门控函数可以区分出各个专家模型层所擅长的领域，从而取得更好的训练效果。

2.3.2 可视化分析

如图 6 所示，通过对不同类型的异常进行测试，我们的方法相较于 Sultani 等人的方法在多个类型上都取

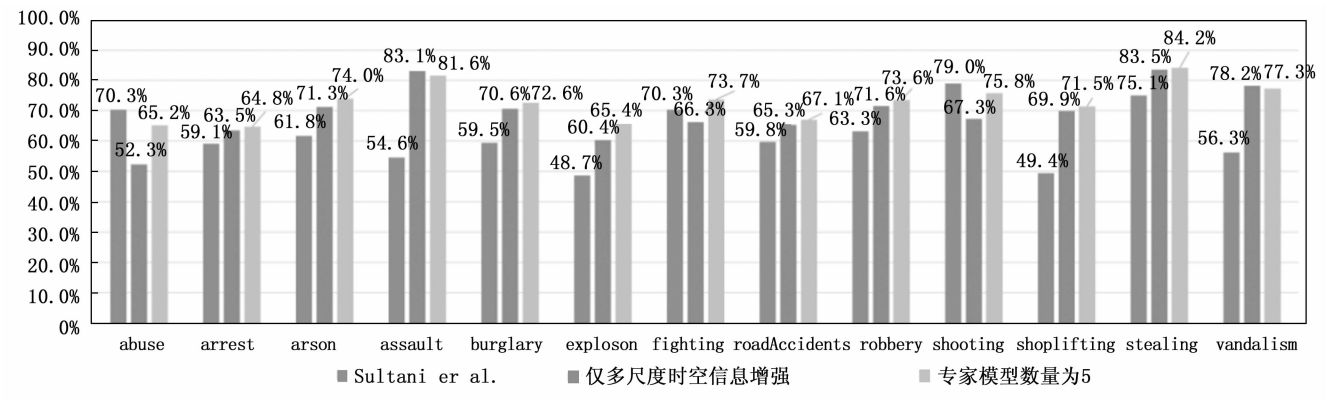

图 6　各类型下检测效果的对比

得了显著的性能提升。特别是在 assault（袭击）、explosion（爆炸）和 shoplifting（店铺抢劫）等基线方法中表现较差的异常类型上，我们的方法表现出了更好的效果。同时，通过比较在不同异常类型上我们的方法与文献［7］的方法的方差，我们的方法显示出了更好的稳定性。这表明我们的方法在处理不同类型异常时能够保持更为一致的性能表现，而不容易受到异常类型变化的影响。

其次，我们的方法由于采用了融合多尺度时空特征信息的策略，对于较为快速、不明显的异常有更好的理解能力。它能够结合 I3D 产生的视频特征，考虑视频内容的前后关系和多方向内容，从而对异常更为敏感。这在多类型的检测效果增加以及对不易察觉的异常类型如殴打的识别率方面得到了体现。

除此之外，我们观察到基于专家模型数量为 5 的异常检测模型由于参考了多个专家模型的得分结果，在多个类型上都取得了显著的性能提升，进一步验证了我们的方法的有效性。

最后，3 个模型的检测准确率和方差如表 4 所示，基于混合专家机制的模型具有更小的准确率方差，相较于融合多尺度时空信息增强的方法减少了 50%。这进一步验证了多专家模型层共同参与视频异常得分，并在各自擅长领域给出更高得分的机制，为模型带来了良好的泛化能力。在时空信息增强中，一些类别的性能下降，比如打枪这一类别相较于基线方法有了更好的检测效果的提升，但在攻击类别中却有下降。这说明多专家机制的均衡影响并未完全被消除，仍对检测结果造成一定影响。然而，总体而言，引入混合专家机制的模型在不同异常类型上表现出更好的鲁棒性和泛化能力。

表 4　3 类模型准确率及在各类异常准确率方差对比

模型	各类准确率方差	AUC/%
Sultani et al.	0.008 69	77.92
融合多尺度时空信息增强的模型	0.007 61	84.02
基于混合专家机制的模型($K=5$)	0.003 75	85.01

图 7 展示了两个模型在 UCF-Crime 数据集中的故

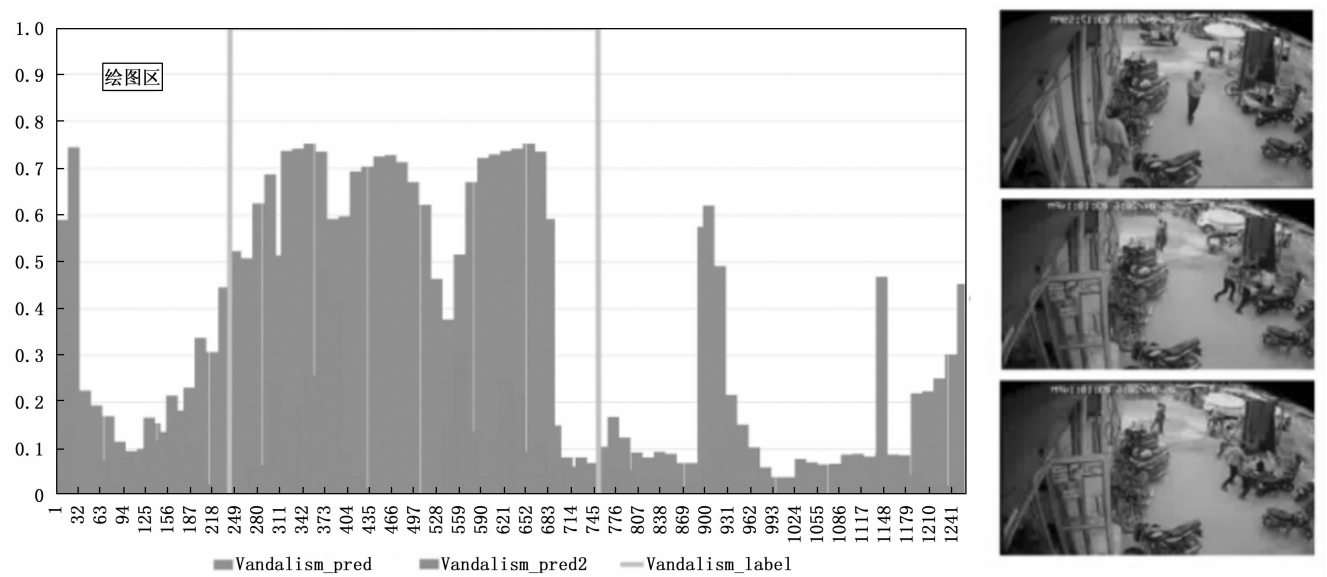

图 7　两种模型在故意破坏类型上的检测效果对比

意破坏类型下的检测表现。图中 x 轴表示视频帧时间定位，y 轴表示对应时间帧的异常得分。图中线段代表人工标注的视频异常得分，其中值为 1 表示对应视频帧存在故意破坏类型异常。深灰色部分为基于混合专家机制的模型，浅灰色部分为 Sultani 等人的方法预测值。

从图中可以观察到，基于混合专家机制的模型在异常片段拥有更高的异常得分，即该模型对异常更为敏感。相比之下，Sultani 等人的方法对故意破坏类型的异常识别能力不足，不能准确识别出异常的存在。这进一步说明了基于混合专家机制的模型在泛化能力上优于 Sultani 等人的方法，从而验证了模型的有效性。

3 结束语

本研究引入了一种名为多尺度时空混合专家增强（MSMOE）的弱监督视频异常检测方法。该方法通过利用多尺度的时间和空间特征进行学习，同时采用了基于混合专家思想的多模型框架，以更好地适应复杂场景下的异常检测任务。通过综合利用这些方法，所建立的模型具备了鲁棒性和区分性。我们在两个具有挑战性的数据集上进行了一系列实验，验证了本文提出的视频异常检测方法在精准性和有效性上的表现，并进一步证明了在复杂场景中采用多专家模型的优越性。这一研究为视频异常检测领域的发展提供了有价值的参考和贡献。

参考文献：

[1] 何 平，李 刚，李慧斌. 基于深度学习的视频异常检测方法综述 [J]. 计算机工程与科学，2022，44 (9)：1620 - 1629.

[2] 付孟丹，宣士斌，王 婷，等. 基于时间和空间注意力机制的视频异常检测 [J]. 计算机技术与发展，2023，33 (8)：51 - 58.

[3] 黄 鑫，肖世德，宋 波. 监控视频中的车辆异常行为检测 [J]. 计算机系统应用，2018，27 (2)：125 - 131.

[4] 王志国，章毓晋. 监控视频异常检测：综述 [J]. 清华大学学报 (自然科学版)，2020，60 (6)：518 - 529.

[5] 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.

[6] GEORGESCU M I，BARBALAUA，LONESCU R T，et al. Anomaly detection in video via self-supervised and multi-task learning [C] // Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington，DC：IEEE Computer Society，2021：12742 - 12752.

[7] SULTANI W，CHEN C，SHAH M. Real-world anomaly detection in surveillance videos [C] //IEEE Conference on Computer Vision and Pattern Recognition，2018：6479 - 6488.

[8] 蔡瑞初，谢伟浩，郝志峰，等. 基于多尺度时间递归神经网络的人群异常检测 [J]. 软件学报，206，26 (11)：2884 - 2896.

[9] 王思齐，胡婧韬，余 广，等. 智能视频异常事件检测方法综述 [J]. 计算机工程与科学，2020，42 (8)：1393 - 1405.

[10] 闫善武，肖洪兵，王 瑜，等. 融合行人时空信息的视频异常检测 [J]. 图学学报，2023，44 (1)：95 - 103.

[11] ZHANG J，QING L，MIAO J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection [C] //2019 IEEE International Conference on Image Processing，2019：4030 - 4034.

[12] ZHONG J X，LI N，KONG W，et al. Graph convolutional label noise cleaner：train a plug-and-play action classifier for anomaly detection [C] //IEEE Conference on Computer Vision and Pattern Recognition，2019：1237 - 1246.

[13] FENG J C，HONG F T，ZHENG W S. Mist：Multiple instance self-training framework for video anomaly detection [C] //IEEE Conference on Computer Vision and Pattern Recognition，2021：14009 - 14018.

[14] TIAN Y，PANG G，CHEN Y，et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning [C] //IEEE International Conference on Computer Vision，2021：4975 - 4986.

[15] RUDER S. An overview of multi- task learning in deep neural networks [EB/OL]. [2017 - 06 - 15]. https：//arxiv. org/abs/1706. 05098v1.

[16] CARREIRA J，ZISSERMAN A. Quo vadis，action recognition a new model and the kinetics dataset [C] // IEEE Conference on Computer Vision and Pattern Recognition，2017：6299 - 6308.

[17] DING J，XUE N，LONG Y，et al. Learning RoI transformer for oriented object detection in aerial images [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re-cognition. New York：IEEE，2019：2849 - 2858.

[18] FAWCETT T. An introduction to ROC analysis [J]. Pattern recognition letters，2006，27 (8)：861 - 874.

[19] LU C，SHI J，JIA J. Abnormal event detection at 150 fps in matlab [C] //IEEE Conference on Computer Vision and Pattern Recognition，2013：2720 - 2727.

[20] LUO W X，LIU W，GAO S H. A revisit of sparse coding based anomaly detection in stacked rnn framework [c] // Proceedings of the IEEE International Conference on Computer Vision. 2017：341 - 349.

[21] CHANG S，LI Y，SHEN J S，et al，Contrastive Attention for Video Anomaly Detection [J]. IEEE Transaction on Multimedia，2022，24：4067 - 4076.