基于全局−局部自注意力网络的视频异常检测方法

杨静[1,2]，吴成茂[3]，周流平[1]

（1. 广州铁路职业技术学院信息工程学院，广东 广州 510430；2. 菲律宾圣保罗大学，土格加劳 3500；
3. 西安邮电大学电子工程学院，陕西 西安 710121）

摘 要：为提升视频异常检测精度，提出一种基于全局−局部自注意力网络的视频异常检测方法。首先，融合视频序列与其对应的 RGB 序列凸显物体的运动变化；其次，通过膨胀卷积层捕获视频序列在局部区域的时序相关性，并利用自注意力网络计算视频全局时序的依赖性，同时，依靠增加基础网络 U-Net 的深度并结合相关运动和表征约束对网络模型进行端到端的训练学习，从而提升模型的检测精度和鲁棒性；最后，对公开数据集 UCSD Ped2、CUHK Avenue 和 ShanghaiTech 进行测试并对所得结果进行可视化分析。实验结果表明，所提方法的检测精度 AUC 值分别达到了 97.4%、86.8%和 73.2%，其性能明显优于对比方法。

关键词：视频异常检测；自注意力；预测；重构

中图分类号：TP391.41

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023151

Novel video anomaly detection method based on global-local self-attention network

YANG Jing[1,2]，WU Chengmao[3]，ZHOU Liuping[1]

1. School of Information Engineering, Guang Zhou Railway Ploytechnic, Guangzhou 510430, China
2. St.Paul University Phillippines, Tuguegarao 3500, Philippines
3. School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Abstract: In order to improve the accuracy of video anomaly detection, a novel video anomaly detection method based on global-local self-attention network was proposed. Firstly, the video sequence and the corresponding RGB sequence were fused to highlight the motion change of the object. Secondly, the temporal correlation of the video sequence in the local area was captured by the expansion convolution layer, along with the self-attention network was utilized to compute the global temporal dependencies of the video sequence. Meanwhile, by deepening the basic network U-Net and combining the relevant motion and representation constraints, the network model was trained end-to-end to improve the detection accuracy and robustness of the model. Finally, experiments were carried out on the public data sets UCSD Ped2, CUHK Avenue and ShanghaiTech, as well as the test results were visually analyzed. The experimental results show that the detection accuracy AUC of the proposed method reaches 97.4%, 86.8% and 73.2% respectively, which is obviously better than that of the compared methods.

Keywords: video anomaly detection, self-attention, prediction, reconstruction

0 引言

视频异常检测中的“异常”与“正常”通常是相对立的。一般而言，相比于正常事件，异常事件的类型是不可穷举的，并且不频繁发生，难以收集。因此，视频异常检测不仅在学术界具有非常重要的

收稿日期：2023−05−15；**修回日期**：2023−07−19

基金项目：广东省高校青年创新人才基金资助项目（No.2020KQNCX198）；广州市基础研究计划基础与应用基础研究基金资助项目（No.104267483017）

Foundation Items: The Young Innovative Talents Project of Guangdong Province (No.2020KQNCX198), Basic and Applied Basic Research Project of Guangzhou Basic Research Program (No.104267483017)

研究价值，在工业界也拥有广阔的应用前景[1-2]。近年来，随着视频监控、故障检测系统、智慧交通及智慧城市等的快速发展，视频异常检测变得尤为重要，视频异常检测的问题得到了国内外学者的广泛关注。在视频异常检测中由于场景和任务属性不同，对异常的定义也不尽相同，Saligrama 等[3]对视频异常检测进行了较准确的定义：视频异常可认为是异常外观或异常运动的属性，或是在异常位置或时间出现正常外观或正常运动属性。在异常检测中，正常数据一般遵循目标类分布，异常数据则是分布外或难以获取的样本。视频异常检测的主要任务是检测出不符合预期规律的罕见样本或从未发生过的突发性事件，而对于这些事件的划分并没有明确的界限和标准。具体而言，根据异常检测应用场景的不同，异常类型的界定和划分也会随之改变，如果用分类的方法解决所有异常事件，则工作量将非常大，难以达到良好的性能。因此，对异常事件的准确检测面临各种挑战，具体表现如下。1) 异常事件的划分因场景而异[4-5]，同一行为在一种任务场景中是正常的，但在另一种任务场景中可能会被判定为异常。2) 异常事件的类型是不可穷举的，对异常事件进行人工标注的工作量非常巨大。3) 一些正常事件与异常事件非常接近，使其区分具有很大的难度。

随着深度学习在动作识别[6-8]、跟踪[9]、轨迹预测[10]、目标检测[11-13]等领域取得成功，视频异常检测得到了大力实践与发展[14-19]。近几年关于视频异常检测的研究主要集中于无监督学习，即在训练模型时仅使用正常样本。首先，通过一分类，进行图像重建/预测，或使用其他自监督学习方式对正常样本进行建模；其次，通过识别不同于训练模型的分布来检测异常。在异常检测中，由于异常数据和正常数据分布不均，呈现长尾分布的特点。因此，相比于有监督学习，无监督学习对视频或图像的异常检测更加合理和有效。基于无监督的深度学习方法不仅易于获取训练的正常样本，而且不需要使用真实的异常样本；无监督的学习范式克服了有监督学习中无法预知异常的问题，因此，拥有更强且有效的特征表达能力。

重构误差作为模型重构能力的评估指标，已被广泛应用于异常检测技术领域[20-22]。重构误差的基本假设如下：一方面，由于正常样本更接近正常训练的数据分布，因此重构误差较异常样本会更低；另一方面，对于非正态分布样本，其假设或预期重构误差会更高[15]。通常基于自动编码器的方法使用重构误差作为识别异常的指标。在传统方法中，为了在卷积神经网络中处理视频序列，将每个图像帧视为具有灰度通道的 2D 图像[23]；然后，将这些灰度帧按照时间顺序堆叠在一起，形成一个新的 2D 图像，其中第三维度由这些堆叠的灰度帧组成。通过这样的堆叠方式，模型可以同时对空间和时间信息进行编码并实现重构。

由于长短期记忆（LSTM, long short term memory）网络能够学习数据的长期依赖关系，Medel 等[24]利用卷积长短期记忆网络进行异常检测，并将该问题定义为重构类型。尽管不是完全的自动编码器，但他们的方法使用了编码器–解码器结构，即给定视频帧的输入序列，卷积长短期记忆网络沿着空间和时间维度提取相关特征；最后，经过解码器并计算重构误差。Hasan 等[25]在第三维度通过堆叠视频帧形成时间立方体，保留必要的时间信息，但这样保留下来的时间信息非常有限。为了解决这个问题，Zhao 等[26]提出通过 3D 卷积保留时间信息，并增加数据来改善样本密度，进而提高检测性能。基于以上工作，Gong 等[15]通过实验测试发现，一些异常事件的重构误差和正常事件的重构误差非常接近，主要是因为自动编码器中卷积神经网络较强的泛化能力，使接近正常的异常事件也被重构出来。为了解决这个问题，Gong 等[15]引入了一种能够将编码特征存储到内存中的自动编码器，即编码器不直接将编码反馈到解码器，而是将编码视为查询，该查询预期返回内存中最接近的正常模式，将该模式用于解码。这样，在重构异常的情况下，由于内存中只含有正常的内存项，因此其重构误差会很高。

近年来，注意力模型被广泛应用于自然语言处理、图像和语音等领域，神经网络的可解释性也被引入无监督的异常检测中。Liu 等[27]使用了类似 grad-CAM（gradient-weighted class activation mapping）[28]的方法将基于梯度的注意力机制推广到变分自动编码器（VAE, variational autoencoder）模型。Venkataramanan 等[29]提出了一种带有注意力引导的卷积对抗变分自动编码器，利用隐空间变量保留的空间信息进行异常定位，并且根据文献[27]的思想生成注意力图，期望在训练时，注意力图可覆盖整个正常区域。Kimura 等[30]利用生成对抗网络（GAN, generative adversarial network）中判别器的注意力图来抑制图像背景造成的异常检测干扰，有效提升了

异常检测模型的鲁棒性。

在数据特征提取的过程中，通常使用卷积来对图像的高维特征信息进行提取，然而卷积操作无论在时间还是空间上均为局部操作。若要获取全局的特征关联性和建立长距离的依赖关系就要构建深层的网络卷积，随着网络深度的增加与卷积块的增多，网络训练的难度增大。因此，单纯的卷积操作对图像的全局信息提取存在一定的局限性。而全局−局部自注意力不仅关注图像局部特征的关联性，还关注特征之间长时间的依赖关系。本文拟采用一种编码器−解码器结构的 U-Net，将 RGB 图像与视频序列 2 种模态信息进行混合编码以突显物体的运动变化，两者共享解码器，得到的特征图通过全局−局部注意力网络处理后再反馈给解码器，从而进行视频异常检测。若解码得到的图像与真实图像差异较大，则表明有异常事件发生，反之则为正常。本文主要工作如下。

1) 采用“双编码器−单解码器”的编解码混合结构，充分利用原始视频的多维信息，并通过自注意力模块实现有效的解码，从而使模型能够准确表示和理解视频数据。

2) 使用多源数据作为输入，充分利用运动和外观信息的互补，并综合考虑不同信息源以全面分析视频数据，从而更加准确地识别异常行为。

3) 提出一种基于全局−局部自注意力机制的视频异常检测方法，通过全局−局部自注意力机制综合考虑整体和局部的时序相关性，能够更好地理解视频序列中不同时间尺度的连续性，并保持局部上下文信息的一致性。

4) 对 UCSD Ped2、CUHK Avenue 和 ShanghaiTech 数据集进行测试，实验结果表明，本文方法的检测精度分别达到 97.4%、86.8%和 73.2%，而且与现有方法相比，本文方法明显提升了视频异常检测的能力和鲁棒性，为视频异常检测的深入研究和实际应用提供了一定支撑。

1　相关工作

1.1　异常检测

许多现有工作将异常检测表述为无监督学习问题，在训练时使用正常数据，并通过重构或判别的方式描述模型的正态性。其中，重构模型将正常数据作为输入映射到某个特征空间，再从特征空间将正常数据映射回输入空间，如自动编码器（AE, autoencoder）[31]、稀疏字典[32]和生成模型[33]。判别模型表征正态样本的统计分布并获得正态实例周围的决策边界，例如，马尔可夫随机场（MRF, Markov random field）[20]、动态纹理混合（MDT, mixture of dynamic texture）[34]、高斯回归[35]和一分类问题[36-37]。然而，这些方法对具有复杂分布的高维数据，如图像、视频等的检测效果欠佳。本文拟采用无监督的深度学习方法进行视频异常检测。

1.2　注意力机制

在深度学习中，模型的参数越多所含信息量越丰富，表达能力也越强，但这也会导致信息量过大的问题。通过引入注意力机制，可快速高效地筛选出高价值的特征信息，使检测模型能更准确地聚焦于关键信息，避免无用信息对模型的干扰，从而克服信息量过大的问题，并提高模型对任务处理的效率和准确性。Purwanto 等[38]在低分辨率视频中利用双向自注意力捕捉长期的时间依赖关系，以此进行视频动作识别。Zhou 等[39]通过注意力图来解决异常检测中前景与背景不平衡的问题，通过对前景和背景赋予不同的权重，使模型更注重前景，并对训练数据中的背景进行有效抑制来提升异常检测性能。Hu 等[40]在自动编码器中引入循环注意力机制，并将其构建为一个循环注意力单元，使模型能够在新场景中具有快速适应能力。Yang 等[41]通过将 Swin Transformer 设计为具有双向跳跃连接的 U 型结构的网络，并在跨注意力和时序上采用残差跳跃连接来进一步辅助还原视频中复杂的静态和动态运动目标特征。

1.3　基于重构和预测的方法

预测模型的目的是将未来的输出帧建模为基于过去若干视频帧的函数，如 GAN 生成未来帧。GAN 主要由两部分组成，一是生成器，模拟原始数据分布；二是判别器，给出来自生成器输入的概率。基于 U-Net 在图像到图像转换方面的出色表现，Luo 等[42]利用类似 GAN 的生成器−判别器结构，将其作为网络的生成器来预测未来帧，并通过网络末端的判别器确定预测帧是否异常。通常假设正常事件是可以预测的，而异常事件则无法预测。Park 等[16]提出了一种在 U-Net 结构下，通过编码器−解码器间的记忆模块所记录的各种正常模式，对未来帧进行预测的方法。同时，Yu 等[43]受到在语言学习中完形填空形式的启发，通过时间维度的上下文和模态信息来建立多个模型，分别预测视频中的视频帧

或视频流。鉴于在实际场景中异常的复杂性，Liu 等[44]提出了一个集成光流重构和视频帧预测的混合框架来进行视频异常检测。首先，在自动编码器中使用多层级记忆模块存储光流重构的正常模式，以便在光流重构误差较大时准确地识别异常事件。其次，在重构光流条件下，通过条件变分自动编码器（CVAE, conditional variational autoencoder）捕捉视频帧和光流之间的高相关性，以便预测未来帧。

在目前主流的异常检测工作中，对正常数据的特征进行重构是较常用且直观的方法。Nguyen 等[17]提出了重构和光流预测共享编码器的网络模型，虽然模型充分学习了物体外观和运动信息的对应关系，但由于光流的计算对资源要求高，整个模型的计算成本较高。在无监督深度学习方法中，AE[31]作为异常检测的常用方法，其对高维数据（如图像、视频等）具有很强的建模能力。基于 AE 的方法通常假设能够重构正常样本，而不能重构异常样本。但由于 AE 的泛化能力过于强大，以至于异常样本也能被很好地重构，因此为了降低 AE 中卷积神经网络（CNN, convolutional neural network）的泛化能力，Chang 等[45]构建了一种将空间和时间信息解耦为 2 个子模块的自动编码器结构，两者同时学习时空特征信息，以提高检测性能。Le 等[46]提出了一种基于残差注意力的自动编码器进行视频异常检测，通过在解码器内引入通道注意力机制对未来帧进行有效预测。由于自动编码器在重构时，缺少对图像某些重点区域编码信息的动态掌握，造成重构时视频帧内容的上下文信息缺失，进而导致模型性能下降。为了解决上述问题，本文基于预测的方法进行异常检测，其主要思想是根据先前若干帧的特征变化来预测当前帧，并在测试阶段将预测出的当前帧与对应的真实帧进行对比，如果两者的预测误差较大，则表明存在异常。这样既充分考虑了正常样本的多样性，又抑制了 CNN 强大的泛化能力。

2 视频异常检测

2.1 基本原理

本文通过对未来帧的预测进行无监督的视频异常检测。受到重构方法的启发[15-16,47]，将预测视为使用之前的若干帧或连续视频序列来进行未来视频帧的重构，因此，本文以一种预测的视角对未来帧进行重构，并采用 U-Net[48]为基础网络框架，进行视频异常检测。全局−局部自注意力网络主要由三部分组成：双编码器、全局−局部自注意力模块、解码器。整个网络均采用端到端的方式进行训练，网络的整体框架如图 1 所示。在输入之前，需要进行简单的数据预处理，即生成与原始图像相对应的 RGB 图像。首先，输入 t 帧的视频序列和对应的 RGB 图像，经过编码器编码后，得到 2 个对应的特征图；然后，将特征图通过按位相加进行融合，将融合后的特征图送入全局−局部自注意力模块进行处理；最后，将处理好的特征图反馈到解码器进行解码，从而进行视频异常检测。

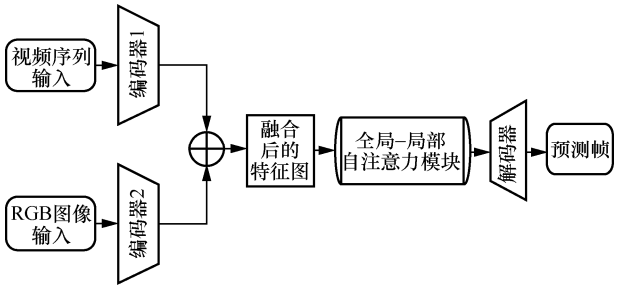


图 1 网络的整体框架

2.2 双编码器−单解码器结构

本文提出的双编码器结构能够很好地对输入图像中的外观和运动信息同时进行学习，并共享一个解码器。本文采用 U-Net 结构，为了避免梯度消失和信息不平衡，U-Net 在高层和低层语义信息之间加入跳跃连接。在原来 U-Net 框架的基础上，本文将网络深度从 4 层增加到 5 层。此外，受 ResNet 结构的启发，本文在模型的主干网络中使用残差模块来代替 U-Net 中的标准卷积模块，但检测效果较差，其原因有两点：其一是 U-Net 整体规模较小，网络没有达到一定深度，使残差模块没有发挥应有的作用；其二，模型训练数据不足，使残差模块得不到充分的训练。

给定编码器 t 帧视频序列 $x_{\text{clips}}=\{I_1,I_2,\cdots,I_t\}$，得到大小为 $H\times W\times C$ 的编码特征图 $\boldsymbol{M}$，其中，H、W和C 分别表示特征的高、宽和通道数。

$$\boldsymbol{M}=f_{\text{e}}(x_{\text{clips}};\theta) \tag{1}$$

其中，θ 为编码器 $f_{\text{e}}(\cdot)$ 的参数。$\boldsymbol{M}$ 经过全局−局部自注意力模块得到特征图 $\boldsymbol{M}'$，并将其反馈到解码器进行解码，即

$$\hat{I}_{t+1} = f_{\mathrm{d}}(\boldsymbol{M}';\alpha) \tag{2}$$

其中，α 为解码器 $f_{\mathrm{d}}(\cdot)$ 的参数。

预测未来帧的损失函数 L_{pre} 和 RGB 损失函数 L_{RGB} 可分别用 L2 损失函数表示为

$$L_{\mathrm{pre}} = \left\| \hat{I}_{t+1} - I_{t+1} \right\| \tag{3}$$

$$L_{\mathrm{RGB}} = \left\| \left| \hat{I}_{t+1} - \hat{I}_t \right| - \left| I_{t+1} - I_t \right| \right\| \tag{4}$$

2.3　全局−局部自注意力模块

根据视频分析和视频理解中注意力机制的相关运行原理[21,49-50]，本文利用全局−局部自注意力模块捕捉时间维度的全局和局部依赖性。膨胀卷积通常应用于空间维度，其主要作用是在同等分辨率的条件下，通过增大卷积的感受野来获得更多的特征信息。本文使用膨胀金字塔卷积，来捕捉视频片段在时间维度上的多尺度依赖性，从而进一步提高视频异常检测性能，全局−局部自注意力框架如图2所示。

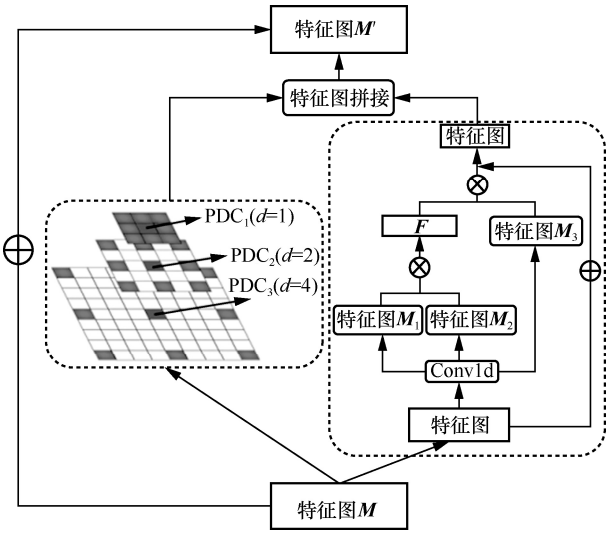


图 2　全局−局部自注意力框架

局部自注意力模块从编码器中得到编码特征图 $\boldsymbol{M} = \{m_1, m_2, \cdots, m_i, m_C\}$，然后对 $\boldsymbol{M}$ 进行卷积操作，在局部自注意力部分主要有三层膨胀卷积操作，即 $\{\mathrm{PDC}_1, \mathrm{PDC}_2, \mathrm{PDC}_3\}$，其对应的膨胀因子 d 分别为{1,2,4}。其数学形式为

$$m^l = \sum_{i=1}^{C} D^l m_i \tag{5}$$

其中，D^l 表示第 l 层的卷积操作，m_i 为特征表达。

全局时序依赖主要通过一个自注意力模块实现，其性能已在视频理解、图像分类、目标检测等多个下游任务中得到验证。通过全局自注意力的作用，将距离相对较远的特征像素点建立一种依赖关系，使全局的特征关联性更加紧密。首先，对特征图 $\boldsymbol{M}$ 进行1×1卷积处理，得到 3 个尺寸和特征相同的特征图 $\boldsymbol{M}_c (c \in \{1,2,3\})$，将特征图 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 的转置进行运算，得到时空关系映射矩阵 $\boldsymbol{F}$，即 $\boldsymbol{F} = (\boldsymbol{M}_1)(\boldsymbol{M}_2)^{\mathrm{T}}$，$F_{ij}$ 表示在位置 i 和位置 j 的关联程度，其数值大小代表了关联性的紧密程度，然后将 $\boldsymbol{F}$ 与 $\boldsymbol{M}_3$ 进行卷积操作，得到 $\boldsymbol{F}' = \mathrm{Conv}_{1\times1}(\boldsymbol{F}\boldsymbol{M}_3)$，将 $\boldsymbol{F}'$ 与原始特征图 $\boldsymbol{M}$ 通过跳跃连接相加得到 $\boldsymbol{F}^{\mathrm{SA}}$，其中 $\boldsymbol{F}^{\mathrm{SA}} = \boldsymbol{F}' + \boldsymbol{M}$。

2.4　损失函数

为了最小化预测帧和真实帧之间的差异，本文使用了强度、梯度和时序图像差异作为约束。强度约束比较两帧之间每个像素的值，保证 RGB 空间的像素值在整个画面中是相似的。梯度约束比较两幅图像相同位置像素值的梯度，并对生成的帧进行锐化。其梯度损失函数为

$$L_{\mathrm{gd}} = \sum_{ij} \left(\left\| \left| \hat{I}_{i,j} - \hat{I}_{i-1,j} \right| - \left| I_{i,j} - I_{i-1,j} \right| \right\|_1 + \left\| \left| \hat{I}_{i,j} - \hat{I}_{i,j-1} \right| - \left| I_{i,j} - I_{i,j-1} \right| \right\|_1 \right) \tag{6}$$

其中，i 和 j 表示像素值的索引位置。在设计梯度损失函数的过程中，本文使用 L1 损失函数作为梯度损失，通常情况下能够得到清晰的图像，并且在训练过程中能够更好地被优化。

对于整个网络模型而言，其整体的损失函数为

$$L = \lambda L_{\mathrm{pre}} + \mu L_{\mathrm{RGB}} + \nu L_{\mathrm{gd}} \tag{7}$$

其中，λ、μ、ν 为超参数。

2.5　异常得分

在最初假设不变的情况下，即模型能够很好地预测正常事件，本文使用预测帧 $\hat{I}$ 与真实帧 I 之间的差异来进行异常预测。均方差（MSE, mean square error）是一种衡量预测图像质量的较常用的方法，其主要思想是通过计算 RGB 图像空间中所有像素的预测值与其真实值之间的欧氏距离。Mathieu 等[51]证实峰值信噪比（PSNR, peak signal to noise ratio）能够很好地对图像质量进行评估，计算式为

$$\mathrm{PSNR}(I, \hat{I}) = 10 \lg \frac{\left[\max_i\right]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i - \hat{I}_i)^2} \tag{8}$$

其中，$\max_i$ 表示图像的最大像素值，$\frac{1}{N}\sum_{i=0}^{N}(I_i-\hat{I}_i)^2$ 表示真实图像与预测图像的像素之间的均方差。PSNR 越高表明该视频帧是正常的可能性就越大，在计算完每帧的 PSNR 之后，将这些数值归一化到[0, 1]内，并计算每个视频帧的异常分数为

$$S(i)=\frac{\mathrm{PSNR}(I_i-\hat{I}_i)-\min_i \mathrm{PSNR}(I_i,\hat{I}_i)}{\max_i \mathrm{PSNR}(I_i,\hat{I}_i)-\min_i \mathrm{PSNR}(I_i,\hat{I}_i)} \quad (9)$$

3 实验结果与分析

本节使用 3 个公开的异常检测数据集测试所提方法以及不同模块的功能，包括 UCSD 行人数据集[34]、CUHK Avenue 数据集[52]和 ShanghaiTech 数据集[53]，并对实验结果进行定性和定量分析，以便验证本文方法的有效性。

3.1 数据集

1) UCSD 行人数据集

UCSD 行人数据集由 Mahadevan 等[34]创建，包含 2 个子数据集 UCSD Ped1 和 UCSD Ped2，该数据集主要通过学校中固定在较高位置的摄像机俯瞰拍摄获得，且人行道的行人密度是由稀疏到稠密不断变化的。UCSD Ped1 中主要包含 34 个训练视频和 36 个测试视频，其分辨率为 238 像素×158 像素。UCSD Ped2 主要包含 16 个训练视频和 12 个测试视频，其分辨率为 360 像素×240 像素。

2) CUHK Avenue 数据集（简称 Avenue 数据集）

CUHK Avenue 数据集[52]采集于香港中文大学（CUHK）校园，数据集中人物的尺寸会因为摄像机的位置和角度而改变。其中共有 47 个异常事件，主要是行人的异常动作及抛物、异常的奔跑等。该数据集包含 16 个训练视频和 21 个测试视频，共 30 652 帧（包括 15 328 个训练帧和 15 324 个测试帧）。

3) ShanghaiTech 数据集

ShanghaiTech 数据集[53]是根据已有数据集的固有缺陷所提出的，即缺乏场景和视角的多样性。数据集包含了 437 个校园监控视频，在 13 个复杂光照条件的应用场景中有 130 个异常视频，由于数据集提出的最初设定是用于无监督学习，因此，异常事件均包含于测试集中。

3.2 评价指标与实验设置

本节实验使用视频异常检测中最常用的评估指标，即接受者操作特征（ROC, receiver operating characteristic）曲线、曲线下面积（AUC, area under curve）和等错误率（EER, equal error rate）。AUC 不关注具体的正负样本得分，只关注整体结果，因此，它能够有效避免在阈值选择过程中因经验设定而产生的主观性，特别适合于正负样本不均衡任务的性能评估。EER 是错误接受率（FAR, false acceptance rate）和错误拒绝率（FRR, false rejection rate）相等时的错误率，也是 ROC 曲线与对角线的交点。模型性能越好，AUC 越高，EER 则相反。根据文献[15,44,47]的实验要求，本文实验使用 NVIDIA GeForce RTX 3090 GPU 进行端到端的训练和测试，网络模型使用 Pytorch 深度学习框架实现，并使用 Adam 随机梯度下降来进行参数优化，学习率为 1×10^{-4}，使用 AUC 对检测模型的性能进行判别。

3.3 方法比较

本节将所提方法与基于手工特征的方法以及基于深度学习的方法进行比较，对比方法如下。1) 基于手工特征的方法：MPPCA[20]、MDT[34]、DFAD[54]。2) 基于深度学习的预测方法：Conv AE[30]、ConvLSTM-AE[55]、TSC[53]、MNAD[16]、IPR[47]等。表 1 列出了不同方法的 AUC，对比方法的性能均是从其对应文献中获得的。

从表 1 可知，所提方法的异常检测精度优于大多数对比方法，在 USCD Ped2、Avenue 和 ShanghaiTech 数据集上的 AUC 分别为 97.4%、86.8%、73.2%，主要得益于其对编码器的特征分别进行了全局和局部的细节处理，使模型性能有了很大的提升。与 IPR[47]相比，本文方法在 3 个数据集上的 AUC 均高出 1%~3%，虽然 IPR 中使用的网络结构也基于编码器−解码器结构，但缺少对物体外观和运动特征等信息的处理；同样地，MNAD[16]也没有对物体外观和运动信息进行有效处理，而本文方法中加入了 RGB 图像的输入，用来增强视频序列的上下文信息，RGB 图像的信息量与光流特征大体相当，但会节省存储空间并加快学习速度，MNAD 中增加了记忆项，存储了丰富的正常事件的原型，使模型在 Avenue 数据集上的性能比本文方法高 1.7%，由此可见，原型学习对无监督视频异常检测任务的研究提供了新的思路，对后续研究有一定的推动作用。与文献[42]相比，本文不仅在模型中加入了运动、外观和上下文信息的相关处理，也在基础网络上增加了网络的深度，使网络的整体性

能有所提升。本文方法与 USTN-DSC[41]都采用了注意力机制，但在 AUC 方面，USTN-DSC 表现出较好的性能，这主要是因为 USTN-DSC 使用了目前最先进的视频处理架构 Swin Transformer，并在时序和注意力中融入了残差连接，能够更好地传递和利用信息，使其性能有了较大提升；此外，HSC[56]采用了一种全新的思路，即引入场景感知的概念进行异常检测，并取得了令人满意的效果，这为解决视频异常检测问题提供了另一种思路和方法。综上分析，在视频异常检测上，本文构建的全局−局部自注意力网络有效性得到了验证。

表 1　　不同方法的 AUC

方法	AUC		
	UCSD Ped2 数据集	Avenue 数据集	ShanghaiTech 数据集
MPPCA[20]	69.3%	—	—
MPPCA+SFA[20]	61.3%	—	—
MDT[34]	82.9%	—	—
DFAD[54]	—	78.3%	—
Conv AE[30]	85.0%	80.0%	60.9%
ConvLSTM-AE[55]	88.1%	77.0%	—
AE-Conv3D[26]	91.2%	77.1%	—
Unmasking[57]	82.2%	80.6%	—
TSC[53]	91.0%	80.6%	67.9%
Stacked RNN[53]	92.2%	81.7%	68%
Frame-Pred[58]	95.4%	84.9%	72.8%
MemAE[15]	94.1%	83.3%	71.2%
AMC[17]	96.2%	86.9%	—
MNAD[16]	97%	88.5%	70.5%
IPR[47]	96.2%	83.7%	71.5%
USTN-DSC[41]	98.1%	89.9%	73.8%
HSC[56]	98.1%	92.4%	83.4%
所提方法	97.4%	86.8%	73.2%

3.4　消融实验分析

本文对模型中所涉及的主要模型组件进行了定量分析，模型组件在 UCSD Ped2 和 Avenue 数据集上性能对比如表 2 所示。增加全局注意力模块后 AUC 仅有小幅提升，在 UCSD Ped2 上 AUC 提升了 0.7%，主要是因为将数据降维编码后，数据的高维特征丢失较多，使全局特征处理受限；而在局部注意力中，现有的编码特征将信息处理的重点放在了细节处理上，使模型性能明显提升，在 UCSD Ped2 上性能提升了 1.6%。实验结果表明，将全局−局部自注意力模块加入模型后在 UCSD Ped2 上的检测效果达到最优，为 97.4%。

表 2　　模型组件在 UCSD Ped2 和 Avenue 数据集上性能对比

模块组件	AUC			
U-Net	√	√	√	√
全局注意力模块	×	√	×	√
局部注意力模块	×	×	√	√
UCSD Ped2	95.2%	95.9%	96.8%	97.4%
Avenue	82.8%	83.3%	85.4%	86.8%

本文在其他实验组件不变的情况下，对模型架构的基础组件在 UCSD Ped2 数据集上进行了测试和性能分析，具体如表 3 所示。通过加深基础主干网络的深度，使网络的非线性表达能力更好，能够学习更复杂的特征变换，从而更好地拟合复杂的特征输入，主干网络的加深使模型检测性能提升了 0.3%。与经典的单编码器−单解码器相比，本文采用的双编码器模式通过加入相比于光流更轻量化的 RGB 图像，将原本单个模态的特征信息转变为 2 种模态信息的有效融合作为输入信息，从而对特征提取起到了增强作用，尤其是对运动信息的加强，使模型性能相较于单编码器结构提升了 0.8%。

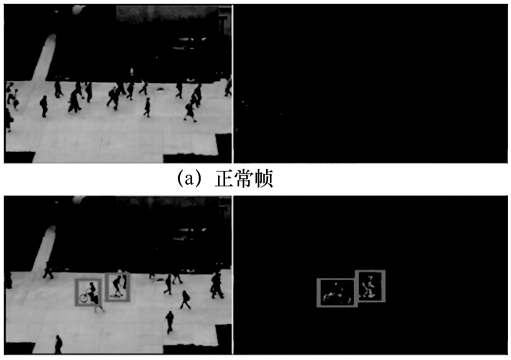
表 3　　模型架构基础组件性能对比

结构组件	AUC
4 层 U-Net	97.1%
5 层 U-Net	97.4% (↑0.3%)
单编码器	96.6%
双编码器	97.4% (↑0.8%)

3.5　可视化分析

本文分别将模型在 UCSD Ped2 和 Avenue 数据集上的测试结果进行了可视化分析。图 3 展示了在 UCSD Ped2 数据集上正常帧和异常帧的检测结果，其中具有异常行为的目标物体已用方框进行了标注，图 3 中的可视化结果主要为了突出显示异常事件发生的位置，将可视化后的原始彩色图转换为黑白图后，正常帧与异常帧的差别非常明显。在正常帧情况下，没有异常发生，此时的异常分值曲线图处于较高位置，对应于图像时，其色彩过度较平缓，被检测物体间的色彩差异大致相同，如图 3(a)所示，在人行横道上的正常情况为正常行走的路人；当有异常发生时，发生异常的位置会显示高异常色彩，如图 3(b)所示，方框标注处为高异常，即有人在人行横道上骑自行车和玩滑板。图 4 展示了 Avenue 数据集测试视频的异常得分。当行人正常行走时，

异常得分处于较高位置，而有人向空中抛掷杂物时，则被判定为一个异常事件，此时异常得分会急剧降低，且异常行为越突出，异常得分越低，这表明本文中的模型能够有效检测到异常事件的发生。



(a) 正常帧

(b) 异常帧

图 3　UCSD Ped2 数据集上正常帧和异常帧的检测结果

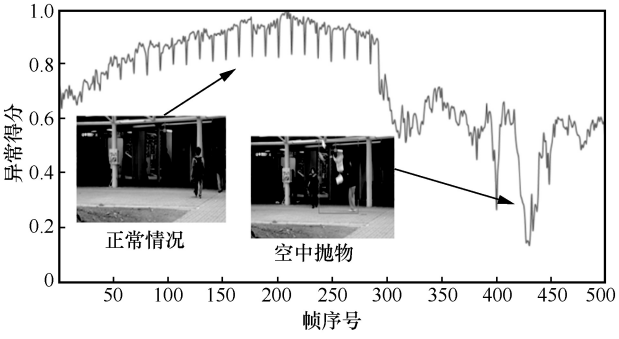


图 4　Avenue 数据集测试视频的异常得分

4　结束语

本文提出了一种基于全局−局部自注意力网络的视频异常检测方法。该方法采用无监督学习方式，通过加深 U-Net 的网络深度、添加多尺度局部注意力模块和全局自注意力模块，以及在数据输入时添加 RGB 图像，增强了模型对视频序列中物体运动、外观等信息的处理能力和鲁棒性。实验结果表明，本文方法在不同应用场景的数据集上具有一定的泛化性和有效性。

CNN 方法通过多层叠加来获得全局信息，但随着叠加层数的增多信息量有所衰减，而 Transformer 中的自注意力机制克服了上述缺陷，使模型具有更强的表达能力，这将是本文未来的研究方向之一。在无监督的方法中，模型的训练通常建立在正常数据集上，如果将已知的异常类型作为重要的先验知识加入模型的训练，则对模型的鲁棒性和检测效果有较大提升。因此，如何将已知的异常类型作为先验知识融入模型的训练将会是本文下一步研究的重点。

参考文献：

[1] RAMACHANDRA B, JONES M J, VATSAVAI R R. A survey of single-scene video anomaly detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2293-2312.

[2] SINGH A, JONES M J, LEARNED-MILLER E G. EVAL: explainable video anomaly localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 18717-18726.

[3] SALIGRAMA V, KONRAD J, JODOIN P M. Video anomaly identification[J]. IEEE Signal Processing Magazine, 2010, 27(5): 18-33.

[4] LUO W X, LIU W, LIAN D Z, et al. Video anomaly detection with sparse coding inspired deep neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 1070-1084.

[5] ŞENGÖNÜL E, SAMET R, ABU A Q, et al. An analysis of artificial intelligence techniques in surveillance video anomaly detection: a comprehensive survey[J]. Applied Sciences, 2023, 13(8): 49-56.

[6] HU T, LONG C, XIAO C. CRD-CGAN: category-consistent and relativistic constraints for diverse text-to-image generation[J]. arXiv Preprint, arXiv: 2107.13516, 2021.

[7] THAKARE K V, RAGHUWANSHI Y, DOGRA D P, et al. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway:IEEE Press, 2023: 5541-5550.

[8] HU T, LONG C J, XIAO C X. A novel visual representation on text using diverse conditional GAN for visual recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 3499-3512.

[9] ISLAM A, LONG C J, BASHARAT A, et al. DOA-GAN: dual-order attentive generative adversarial network for image copy-move forgery detection and localization[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 4675-4684.

[10] GU T P, CHEN G Y, LI J L, et al. Stochastic trajectory prediction via motion indeterminacy diffusion[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 17092-17101.

[11] ISLAM A, LONG C J, RADKE R. A hybrid attention mechanism for weakly-supervised temporal action localization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 1637-1645.

[12] LI Z, WANG Y C, ZHANG N, et al. Deep learning-based object detection techniques for remote sensing images: a survey[J]. Remote Sensing, 2022, 14(10): 2385.

[13] ZHAO Z J, WEI S T, CHEN Q C, et al. Masked retraining teacher-student framework for domain adaptive object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 1-12.

[14] LU Y W, KUMAR K M, NABAVI S S, et al. Future frame prediction using convolutional VRNN for anomaly detection[C]//Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. Piscataway: IEEE Press, 2019: 1-8.

[15] GONG D, LIU L Q, LE V, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anoma-

ly detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 1705-1714.

[16] PARK H, NOH J, HAM B. Learning memory-guided normality for anomaly detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 14360-14369.

[17] NGUYEN T N, MEUNIER J. Anomaly detection in video sequence with appearance-motion correspondence[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 1273-1283.

[18] LIU Z, WU X M, ZHENG D, et al. Generating anomalies for video anomaly detection with prompt-based feature mapping[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 24500-24510.

[19] WANG X, ZHANG S, CEN J, et al. CLIP-guided prototype modulating for few-shot action recognition[J]. arXiv Preprint, arXiv: 2303.02982, 2023.

[20] KIM J, GRAUMAN K. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 2921-2928.

[21] SAUNSHI N. Towards understanding self-supervised representation learning[D]. Princeton: Princeton University, 2022.

[22] WANG Y Z, QIN C, BAI Y, et al. Making reconstruction-based method great again for video anomaly detection[C]//Proceedings of IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2023: 1215-1220.

[23] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

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

[25] HASAN M, CHOI J, NEUMANN J, et al. Learning temporal regularity in video sequences[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 733-742.

[26] ZHAO Y R, DENG B, SHEN C, et al. Spatio-temporal autoencoder for video anomaly detection[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 1933-1941.

[27] LIU W, LI R, ZHENG M, et al. Towards visually explaining variational autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8642-8651.

[28] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 618-626.

[29] VENKATARAMANAN S, PENG K C, SINGH R V, et al. Attention guided anomaly localization in images[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 485-503.

[30] KIMURA D, CHAUDHURY S, NARITA M, et al. Adversarial discriminative attention for robust anomaly detection[C]//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2020: 2161-2170.

[31] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. arXiv Preprint, arXiv: 1312.6114, 2013.

[32] ZHAO B , LI F F ,XING E P . Online detection of unusual events in videos via dynamic sparse coding[C]//Proceedings of Computer Vision & Pattern Recognition. Piscataway: IEEE Press, 2011:3313-3320.

[33] VASWANI N, ROY-CHOWDHURY A K, CHELLAPPA R. “Shape Activity”: a continuous-state HMM for moving/deforming shapes with application to abnormal activity detection[J]. IEEE Transactions on Image Processing, 2005, 14(10): 1603-1616.

[34] MAHADEVAN V, LI W X, BHALODIA V, et al. Anomaly detection in crowded scenes[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2010: 1975-1981.

[35] CHENG K W, CHEN Y T, FANG W H. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 2909-2917.

[36] RUFF L, VANDERMEULEN R A, GÖRNITZ N, et al. Deep one-class classification[C]//Proceedings of International Conference on Machine Learning. New York: PMLR, 2018: 4393-4402.

[37] SCHÖLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[38] PURWANTO D, PRAMONO R R A, CHEN Y T, et al. Corrections to “three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos”[J]. IEEE Signal Processing Letters, 2020, 27: 2188.

[39] ZHOU J T, ZHANG L, FANG Z W, et al. Attention-driven loss for anomaly detection in video surveillance[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(12): 4639-4647.

[40] HU C, WU F, WU W, et al. Normal learning in videos with attention prototype network[J]. arXiv Preprint, arXiv: 2108.11055, 2021.

[41] YANG Z, LIU J, WU Z, et al. Video event restoration based on keyframes for video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 14592-14601.

[42] LUO W X, LIU W, LIAN D Z, et al. Future frame prediction network for video anomaly detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7505-7520.

[43] 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. New York: ACM Press, 2020: 583-591.

[44] LIU Z A, NIE Y W, LONG C J, et al. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 13568-13577.

[45] CHANG Y, TU Z, XIE W, et al. Video anomaly detection with spatio-temporal dissociation[J]. Pattern Recognition, 2022, 122: 108213.

[46] LE V T, KIM Y G. Attention-based residual autoencoder for video anomaly detection[J]. Applied Intelligence, 2023, 53(3): 3240-3254.

[47] TANG Y, ZHAO L, ZHANG S, et al. Integrating prediction and recon-

struction for anomaly detection[J]. Pattern Recognition Letters, 2020, 129: 123-130.

[48] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[49] LIU C Y, XU X Y, ZHANG Y J. Temporal attention network for action proposal[C]//Proceedings of 2018 25th IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2018: 2281-2285.

[50] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7794-7803.

[51] MATHIEU M, COUPRIE C, LECUN Y. Deep multi-scale video prediction beyond mean square error[J]. arXiv Preprint, arXiv: 1511.05440, 2015.

[52] LU C W, SHI J P, JIA J Y. Abnormal event detection at 150 FPS in MATLAB[C]//Proceedings of 2013 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2014: 2720-2727.

[53] LUO W X, LIU W, GAO S H. A revisit of sparse coding based anomaly detection in stacked RNN framework[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 341-349.

[54] GIORNO A D, BAGNELL J A, HEBERT M. A discriminative framework for anomaly detection in large videos[C]//European Conference on Computer Vision. Cham: Springer, 2016: 334-349.

[55] LUO W X, LIU W, GAO S H. Remembering history with convolutional LSTM for anomaly detection[C]//Proceedings of 2017 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2017: 439-444.

[56] SUN S, GONG X. Hierarchical semantic contrast for scene-aware video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 22846-22856.

[57] IONESCU R T, SMEUREANU S, ALEXE B, et al. Unmasking the abnormal events in video[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2914-2922.

[58] LIU W, LUO W X, LIAN D Z, et al. Future frame prediction for anomaly detection - A new baseline[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6536-6545.

[作者简介]

杨静（1986− ），女，陕西西安人，圣保罗大学博士生，广州铁路职业技术学院讲师，主要研究方向为智能视觉识别等。

吴成茂（1968− ），男，四川仪陇人，西安邮电大学高级工程师，主要研究方向为智能信息处理、非线性动力系统与混沌、信息安全等。

周流平（1973− ），男，湖南醴陵人，广州铁路职业技术学院高级工程师，主要研究方向为通信技术、智能信息处理等。