智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于空时对抗变分自编码器的人群异常行为检测

邢天祎, 郭茂祖, 陈加栋, 赵玲玲, 陈琳鑫, 田乐

引用本文:
邢天,郭茂祖,陈加栋,赵玲玲,陈琳鑫,田乐. 基于空时对抗变分自编码器的人群异常行为检测[J]. 智能系统学报, 2023, 18(5): 994–1004.
XING Tianyi, GUO Maozu, CHEN Jiadong, et al. Detection of abnormal crowd behavior based on spatial–temporal adversarial variational autoencoder[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(5): 994–1004.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202303002

DOI: 10.11992/tis.202303002
网络出版地址: https://kns.cnki.net/kcms2/detail/23.1538.TP.20230615.1443.008.html

基于空时对抗变分自编码器的人群异常行为检测

邢天祎[1], 郭茂祖[1], 陈加栋[1], 赵玲玲[2], 陈琳鑫[2], 田乐[1]
(1. 北京建筑大学 电气与信息工程学院, 北京 100044; 2. 哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 150001)

摘 要: 基于视频的人群异常行为检测对提前发现安全风险、预防群体安全事故发生具有重要价值。针对人群异常行为事件的稀少性导致的无法直接充分学习异常样本的表示、异常事件检测精度低的问题, 在变分自编码器基础上, 提出一种基于预测的空时对抗变分自编码器(spatial-temporal adversarial variational autoencoder, ST-AVAE)视频异常检测模型, 通过引入长短期记忆网络(long short-term memory, LSTM)和对抗网络模块, 对正常样本视频序列的时间维度与空间维度进行联合特征表示与重构, 减少了正常样本重建过程中的特征损失进而扩大了异常样本的预测损失, 避免了对异常样本的依赖, 实现了基于模型重构误差的人群逃散异常行为检测。在公开数据集 UMN 及采集视频数据集上进行对比实验, 证明 ST-AVAE 模型在基于监控视频的人群异常逃散行为检测中均具有最优的检测精度和召回率, 对抗网络模块显著提升了异常检测的性能。

关键词: 人群异常行为检测; 变分自编码器; 自编码器; 长短期记忆网络; 对抗网络; 空时对抗变分自编码器; 重构误差; 异常逃散行为
中图分类号: TP181 **文献标志码**: A **文章编号**: 1673−4785(2023)05−0994−11

中文引用格式: 邢天祎, 郭茂祖, 陈加栋, 等. 基于空时对抗变分自编码器的人群异常行为检测 [J]. 智能系统学报, 2023, 18(5): 994–1004.
英文引用格式: XING Tianyi, GUO Maozu, CHEN Jiadong, et al. Detection of abnormal crowd behavior based on spatial-temporal adversarial variational autoencoder[J]. CAAI transactions on intelligent systems, 2023, 18(5): 994–1004.

Detection of abnormal crowd behavior based on spatial-temporal adversarial variational autoencoder

XING Tianyi[1], GUO Maozu[1], CHEN Jiadong[1], ZHAO Lingling[2], CHEN Linxin[2], TIAN Le[1]
(1. School of Electrical and Information Engineering, Beijing, University of Civil Engineering and Architecture, Beijing 100044 China; 2. Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Video-based detection of abnormal crowd behavior is important for the early discovery of safety risks and the prevention of group safety accidents. To address insufficient direct learning of the representation of abnormal samples because of the scarcity of abnormal crowd behavior events and the low detection accuracy of abnormal events, this study proposed a predictive spatiotemporal adversarial variational autoencoder (ST-AVAE) video anomaly detection model based on variational autoencoder, by adding the long short-term memory and adversarial network modules. Joint feature representation and reconstruction of the temporal and spatial dimensions of normal sample video sequences were performed, which reduced the feature loss in the reconstruction process of normal samples, thereby expanding the prediction loss of abnormal samples, avoiding dependence on abnormal samples, and realizing the detection of the abnormal behavior of crowd dispersal based on model reconstruction errors. Comparative experiments were conducted on the public dataset UMN and captured video datasets to prove that the ST-AVAE model has the optimal detection accuracy and recall rate in the detection of abnormal crowd escape behavior based on surveillance video, and the adversarial network module significantly improves the performance of anomaly detection.

Keywords: detection on abnormal crowd behavior; variational autoencoder; autoencoder; long short-term memory network; adversarial network; spatiotemporal adversarial variational autoencoder; reconstruction errors; abnormal escape behavior

视频异常检测指基于视频数据检测其中不符合正常预期的行为、事件等[1]。随着监控设备的广泛普及与计算机视觉技术的快速发展, 基于视频的异常检测技术被广泛应用于交通管控、智慧

收稿日期: 2023−03−01. **网络出版日期**: 2023−06−16.
基金项目: 国家自然科学基金项目(62271036, 61871020); 国家重点研发计划项目(2021YFF0306303); 北京市属高校高水平创新团队建设计划项目(IDHT20190506).
通信作者: 赵玲玲. E-mail: Zhaoll@hit.edu.cn.

安防、事故预警等诸多领域, 为大量实际应用场景提供了支撑。在踩踏、挤压等群体事故形成初期通常伴随有群体异常动向[2], 通过检测监控视频中的人群异常行为, 有助于及时感知事故危险隐患, 对提升公共安全监管效率、避免重大群体事件具有重要的研究意义与研究价值。

目前基于深度学习的方法越来越多地应用于视频异常行为检测, 这类方法通过自动的从大量数据集中学习数据本身的分布规律来提取出更加鲁棒的高级特征, 具有更强的特征表示能力。目前, 基于深度学习的视频异常行为检测方法主要分为基于重构和基于预测两类。

基于重构误差的方法是通过模型训练学习正常样本在样本空间服从的分布, 符合该分布的正常样本都能较好地重构, 而那些重构误差大的样本则属于异常样本。Hasan 等[3] 利用 2D 卷积自动编码器 (two dimensional-convolutional autoencoder, 2D-CAE) 来重构正常帧并使用多个帧作为输入, 但所提出的网络仅在空间上执行卷积和池化运算, 无法从视频中捕获时间模式。因此文献 [4-6] 通过利用卷积长短期记忆自编码器 (convolution long-short term memory autoencoder, Conv LSTM-AE) 重构目标对象的外观信息和运动信息进行异常行为检测, 提出将稀疏编码映射到堆叠的循环神经网络 (stacked recurrent neural network, sRNN) 框架中重构异常行为。但由于卷积神经网络具有的强大的泛化能力, 某些异常事件的重构误差也较小。Yan 等[7] 提出了双流循环变分自编码器模型 (two-stream recurrent variational autoencoder), 双流融合架构在异常事件检测中用于融合空间流和时间流的信息, 实现了异常事件的帧级检测及像素级定位。Liu 等[8] 提出了双原型自编码器 (dual prototype autoencoder, DPAE), 引入了双原型损失和重构损失, 使编码器产生的潜在向量更接近自己的原型, 因此潜在向量趋于接近, 则表示正常, 潜在向量距离较大则表示异常。但是此类方法均受限于数据样本不均衡, 正常样本重构误差占主导地位等问题, 在某些场景下不能准确检测出异常事件。

基于预测的视频异常检测方法假设正常行为是有规律的且是可预测的, 而视频中异常行为事件由于其不确定性不可预测。该类方法可通过生成未来目标帧的预测帧, 将其与对应的视频真实帧进行对比来判断该视频中是否包含异常行为。目前, 生成对抗网络 (generative adversarial network, GAN) 在视频异常检测领域已取得突破性进展, 其网络架构可很好地用于预测。Liu 等[9] 提出基于 U-net 的条件生成对抗网络进行异常行为检测, 并采用 Flownet 光流网络对运动特征约束; Dong 等[10] 在此基础上提出基于对偶生成对抗网络模型, 利用双生成器和双辨别器的对偶结构分别对外观和运动信息判断异常。Nguyen 等[11] 采用卷积自编码器网络学习空间维度特征, 与运动信息相关联输入 U-net 网络实现异常检测。通过向传统卷积自编码器引入 GAN 的辨别器结构, 文献 [12] 构建了对抗自编码器 (adversarial autoencoder, AAE) 模型, 该对抗式自编码器由传统的卷积自编码器 (convolutional autoencoder, CAE)[5] 和辨别器[13] 组成, 使输入样本和输入潜在表示与重构样本与输出的潜在表示之间分别形成对抗关系。Li 等[14] 在对抗式自编码器的基础上提出空时对抗自编码器 (spatial-temporal adversarial autoencoder, ST-AAE) 模型, 基于视频数据的空时特征进行预测, 实现了异常行为的检测功能。Zhang 等[15] 提出了一种融合变分自编码 (variational auto-encoder, VAE) 和分阶段生成对抗网络 (stack generative adversarial networks, StackGAN) 的生成模型, 进一步提高了生成图像的质量。但是这些方法较多针对个体行为异常检测, 对群体行为异常的研究仍不充分。

在最近的研究中, Park 等[16] 提出使用基于 CNN 的记忆引导法异常检测 (memory-guided normality for anomaly detection, MNAD) 对视频数据进行异常检测。Markovitz 等[17] 提出了时空图自编码 (spatio-temporal graph autoencoder, ST-GCAE) 来检测异常人体姿势。Goyal 等[18] 提出了一种用于无监督异常检测的深度鲁棒单类分类 (deep robust one-class classifica, DROCC)。他们的方法假设来自正常类的点位于一个良好采样和局部线性低维流形上, 通过学习一个表示来最小化分类损失, 然后使用分类器将正常样本从异常样本中分离出来。为了构建一个高性能的缺陷检测模型, 能够从没有异常数据的图像中检测出未知的异常模式, Li 等[19] 提出了一种用于构建异常检测器的两阶段 CNN, 通过数据增强策略 (CutPaste) 对正常数据进行分类来学习表示。Rudolph 等[20] 提出的 CS-Flow (cross-scale-flows) 用一种新颖的全卷积跨尺度归一化流, 该流联合处理不同尺度的多个特征映射。该方法保持了空间排列, 使得归一化流的潜在空间是可解释的, 这使得该方法能够定位图像中的缺陷区域。Carrara 等[21] 提出了基于双头对抗生成网络的 CBiGAN(consistency bidirectional generative adversarial network), 用 GAN 和 AutoEncoder 的结合来学习正常数据的分布, 然后通过重构误差来判断当前图像是否异常。但此类方法在重建图像上能力较差, 导致正常样本重构误差较大, 异常事件检测精度较低。

目前, 结合对抗自编码器结构与空时特征的

视频异常行为检测方法已取得了较好的效果，但仍存在部分局限性：1）现有研究较多针对个体或局部异常行为进行检测，对群体异常行为的研究仍不充分；2）视频数据由单帧图像组成，现有方法主要采用图像检测的方法进行视频异常检测，损失时序信息；3）召回率低，由于异常事件罕见且具有差异性，识别所有的异常较为困难，导致正常样本被误报为异常，真实且复杂的异常却被漏报。

为解决当前研究存在的问题，本文提出了一种基于重构和预测相结合的异常检测模型：空时对抗变分自编码器（spatio-temporal adversarial variational autoencoder, ST-AVAE）。模型同时融合了长短时记忆网络，变分自编码器模块以及对抗网络模块。保留了视频数据的时序信息，在变分自编码器生成重构帧图像时，加入了对抗网络模块，进一步提高了重构图像的能力，降低了正常样本重构误差，增大了异常样本重构误差，进而提升异常事件检测精度。

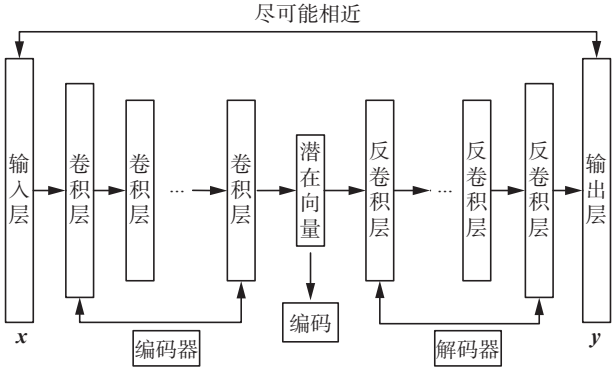
本文的主要创新在 ST-VAE 模型基础上提出 ST-AVAE 模型，将 GAN 模型的辨别器与 ST-VAE 结合，辨别器旨在使 ST-VAE 模型学习到模拟正常数据分布的能力，提高对正常样本空时特征的表示和重建能力，同时辨别器的引入使得异常样本和正常样本的表示区分度更强，从而重建误差具有显著不同，提高对异常的检测能力。

1　相关工作

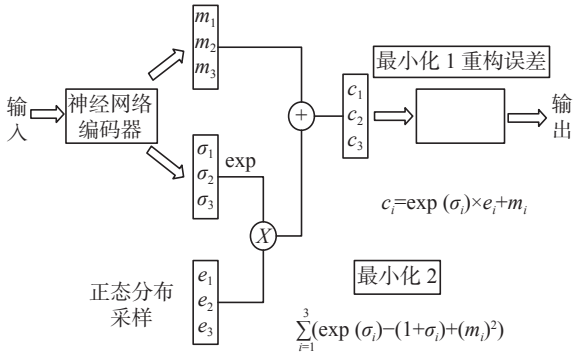
本文在解决人群异常检测正负样本不均衡，过于依赖异常样本的问题上，采用了变分自编码器作为模型基础，并结合了长短时记忆网络，提取了视频数据的时序信息。

1.1　变分自编码器

自编码器（autoencoder, AE）由编码器–解码器（encoder-decoder）组成，通过将输入信息作为学习目标，对输入信息进行表征学习。如图 1(a) 所示，输入原图像数据 x，通过多层卷积层得到潜在向量，再经多层反卷积层得到生成图像 $\boldsymbol{y}$，模型训练过程旨在使 $\boldsymbol{y}$ 尽可能与 $\boldsymbol{x}$ 相似。



(a)自编码器网络示意



(b) 变分自编码器网络示意

图 1　自编码器网络示意

Fig. 1　AutoEncoder network

变分自编码器（variational autoencoder, VAE）在自编码器模型上做进一步变分处理，使得编码器的输出结果能对应到目标分布的均值和方差。如图 1(b) 所示，VAE 在生成潜在向量 (c_1,c_2,c_3) 前，会向编码添加噪音以加大潜在向量空间，编码器输出两个编码，一个是原有编码 (m_1,m_2,m_3)，另一个是控制噪音干扰程度的编码 $(\sigma_1,\sigma_2,\sigma_3)$，第 2 个编码为随机噪音码分配权重 (e_1,e_2,e_3)，通过 $\exp(\sigma_i)$ 保证这个分配权重为正，最后将原编码与噪音编码相加，即可得到 VAE 在 code 层的输出结果。

1.2　长短时记忆网络

长短时记忆（long short-term memory, LSTM）网络是一种时间循环神经网络，解决一般循环神经网络（recurrent neural network, RNN）存在的长期依赖问题。LSTM 的主要作用是舍去重要性较低的信息，并将较为关键的信息随时间传递到下一时刻，由此达到预测目的。LSTM 单元结构中包含 3 种门控机制，输入门、遗忘门、输出门，如图 2 所示。

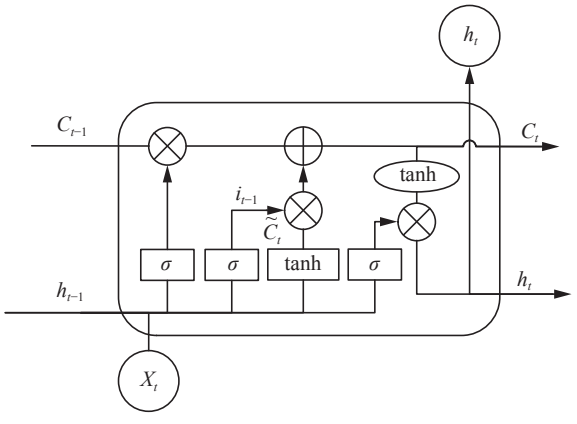


图 2　LSTM 结构

Fig. 2　LSTM structure diagram

图 2 中 h_{t-1},C_{t-1} 分别代表 LSTM 上一单元的输出和状态, X_t, h_t, C_t 分别代表当前时刻输入、输出和状态。状态 C_{t-1} 会被上一时刻输出 h_{t-1} 及当前时刻输入 X_t 通过 3 种门结构进行计算, 得到当前时刻状态 C_t, 当前时刻输出 h_t 以同样形式参与下一时刻状态计算。图中 σ 代表 sigmoid 激活函数, tanh 代表双曲正切激活函数。

首先, 遗忘门决定上一时刻状态 C_{t-1} 中保留和删除的信息, 其公式为

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

f_t 将与 C_{t-1}相乘, 由于 σ 函数取值 0~1, C_{t-1} 与 0 相乘的位置信息将被遗忘。

输入门决定新输入带来的信息, 计算过程为

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$C_t' = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * C_t' \tag{4}$$

输出门决定最后需要输出的信息, 计算过程如下:

$$O_t = \sigma(W_0 \cdot [h_{t-1}, x_t] + b_0) \tag{5}$$

$$H_t = o_t * \tanh(C_t) \tag{6}$$

2　人群异常行为检测算法模型

2.1　模型整体框架

针对监控场景下人群异常行为检测问题, 本文利用 ST-VAE 和 GAN 网络的辨别器结构, 设计了空时对抗变分自编码器模型, 以提高异常行为检测能力。模型由 CNN 残差网络构成的编码器、LSTM 组成的空时预测模块、解码器和辨别器 4 部分组成。在编码器部分, 输入视频帧序列 $(x_k, x_{k+1}, \cdots, x_{k+m})$, 生成视频帧序列的特征潜在向量 $(e_k, e_{k+1}, \cdots, e_{k+m})$; 在 LSTM 网络层, 对 $(e_k, e_{k+1}, \cdots, e_{k+m})$ 进行预测, 得到预测帧向量 $(e'_k, e'_{k+1}, \cdots, e'_{k+m+1})$; 在解码器部分, 由特征编码重建$k+1, k+2, \cdots, k+m+1$时刻的视频帧序列 $(x'_{k+1}, x'_{k+2}, \cdots, e'_{k+m+1})$, 与真实样本 $(x_k, x_{k+1}, \cdots, x_{k+m})$ 计算重建误差; 最后, 将视频帧序列与生成的潜在向量(真实样本对)和重构视频帧序列与预测潜在向量(生成样本对)输入到对抗自编码器的辨别器结构, 对原帧评价分、重构帧评价分和重建误差加权求和后, 与选定的最佳阈值进行比较, 判定是否发生人群异常行为。空时对抗变分自编码器网络(ST-AVAE)的整体结构如图 3 所示。

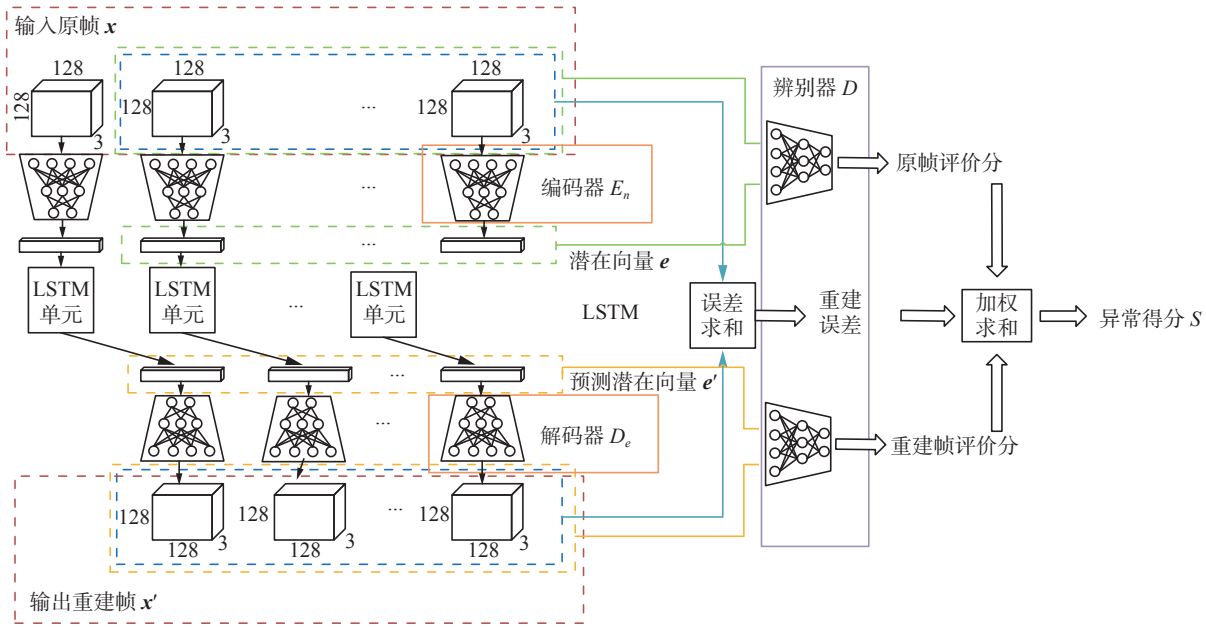


图 3　ST-AVAE 模型结构

Fig. 3　ST-AVAE model structure

2.2　编码器模块

本文采用的编码器结构如图 4 所示, 图中所有代表卷积层的蓝色长方块部分, 均由残差模块[22]代替。编码器部分由残差模块和平均池化层组成, 输入为 128×3×3 尺寸的图像数据, 输入经 6 层卷积层, 1 层全连接层, 变为 256×1 维的向量, 再通过 Leaky-ReLU 激活层将向量分为两个 64×1 维的向量, 分别代表均值和方差, 得到一个近似正态分布的潜在向量 $\boldsymbol{e}$。

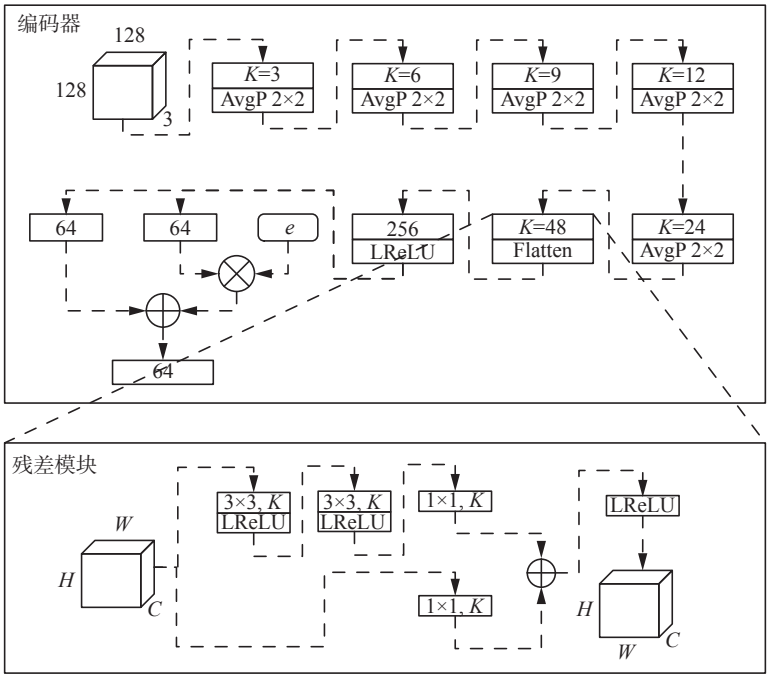


图 4　编码器模块

Fig. 4　Encoder module

编码器的工作原理可表示为

$$\boldsymbol{e}_i = E_n(\boldsymbol{x}_i) \tag{7}$$

其中，输入 $\boldsymbol{x}_i$ 表示第 i 帧原图，经过编码器 E_n 得到第 i 帧编码向量 $\boldsymbol{e}_i$。

2.3　空时预测模块

在 2.2 节中编码器得到的隐变量上增加 LSTM 模块，结构如 1.2 节图 2 所示。空时预测模块的输入为前 $k-1$ 帧序列得到的潜在向量，得到 2~k 帧的预测潜在向量，即：

$$\boldsymbol{e}_i' = \mathrm{LSTM}(e_i) \tag{8}$$

$\boldsymbol{e}_i=[e_1\ e_2\ \cdots\ e_k]$表示 k 帧序列通过编码得到的 k 个潜在向量，$\boldsymbol{e}_i'=[e_2'\ e_3'\ \cdots\ e_k']$表示通过 LSTM 单元后得到的预测帧序列潜在向量。

2.4　解码器模块

解码器由残差模块和上采样层组成。解码器部分通过进行尺寸与编码器对应的反卷积层和上采样层，将潜在向量解码成 128×3×3 与原图相同大小的生成图像。通过 Decoder 层解码回 2~n 帧的重构帧序列。解码器结果如图 5 所示。

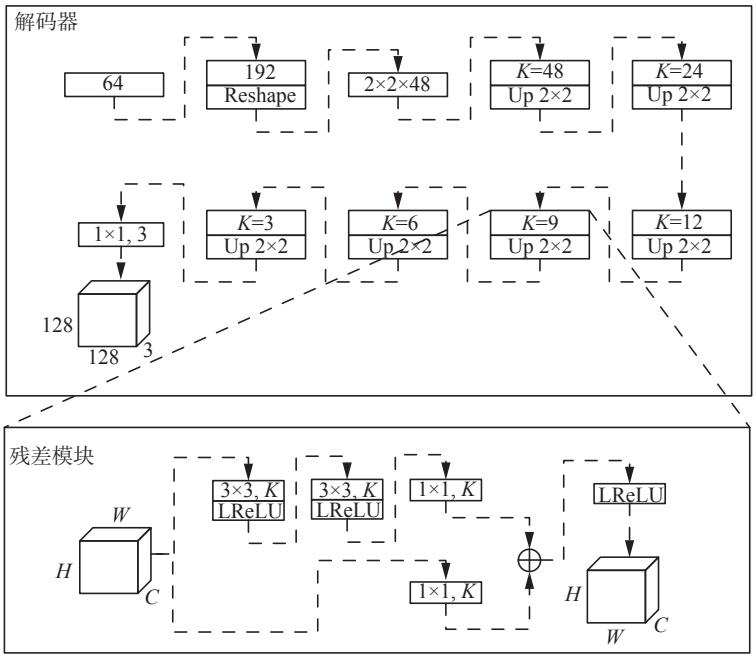


图 5　解码器模块

Fig. 5　Decoder module

解码器的工作原理可由下式表示:

$$\boldsymbol{x}_i' = D_e(\boldsymbol{e}_i') \tag{9}$$

其中, 输入$\boldsymbol{e}_i'$表示输入第 i 帧预测的潜在向量, 经过解码器模块得到第 i 帧重构帧$\boldsymbol{x}_i'$。

2.5　辨别器与对抗学习模块

为了使 VAE 模型更好地学习到模拟正常数据分布的能力, 提高模型的泛化能力, 因此在模型中加入了辨别器, 利用对抗学习的方式来强化编码器–解码器的重构图像能力。

辨别器–编码器–解码器共同形成对抗网络, 整个对抗网络首先更新其辨别器以区分真实样本(服从正态分布)和生成样本(由编码器计算得到的潜在向量), 然后更新其生成器(编码器–解码器)以混淆辨别器。辨别器结构如图 6 所示, 其目标是尽量使生成的虚假图片和隐藏层向量对(即重建帧和预测潜在向量)与真实图片和生成的隐藏层向量对(即原帧和潜在向量)尽量无法区分哪对才是正常样本对。

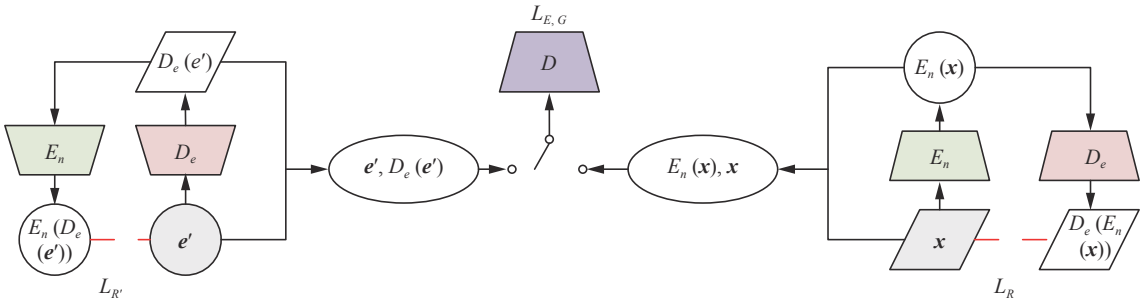


图 6　辨别器与对抗学习模块

Fig. 6　Discriminator and adversarial learning module

图 6 左半部分输入为预测潜在向量$\boldsymbol{e}_i'$, 通过解码器网络 D_e 生成图像, 然后再用编码器网络 E_n 映射成潜在向量, 求重建误差:

$$L_{R'}(\boldsymbol{e}') = \|\boldsymbol{e}' - E_n(D_e(\boldsymbol{e}'))\|_1 \tag{10}$$

图 6 右半部分输入为原帧图像 $\boldsymbol{x}$, 用编码器网络 E_n 映射生成潜在向量, 用网络 D_e 映射生成重构图像, 求重建误差:

$$L_R(\boldsymbol{x}) = \|\boldsymbol{x} - D_e(E_n(\boldsymbol{x}))\|_1 \tag{11}$$

网络 E_n、D_e 的目标首先要使这两个误差尽可能地小:

$$L_c(\boldsymbol{x},\boldsymbol{e}') = L_R(\boldsymbol{x}) + L_{R'}(\boldsymbol{e}') \tag{12}$$

式中: $L_c(\boldsymbol{x},\boldsymbol{e}')$ 为网络 E_n、D_e 的重建误差损失函数, 它是两个重建误差的加和。

这里用 GAN 的损失函数为

$$L_D = -\frac{1}{N}\sum_{i=1}^{N} D(\boldsymbol{x}_i, E_n(\boldsymbol{x}_i)) + \frac{1}{N}\sum_{i=1}^{N} D(D_e(\boldsymbol{e}_i'), \boldsymbol{e}_i') \tag{13}$$

即式(7)中 $\boldsymbol{x}_i$ 和 $E_n(\boldsymbol{x}_i)$ 组成的样本对视为正样本对, 式(9)中$\boldsymbol{e}_i'$和 $D_e(\boldsymbol{e}_i')$ 组成的样本对视为负样本对。辨别器试图增大正样本对的评分, 减小负样本对的评分。

$$L_{E_n,D_e} = \frac{1}{N}\sum_{i=1}^{N} D(x_i, E_n(x_i)) - \frac{1}{N}\sum_{i=1}^{N} D(D_e(e_i'), e_i') \tag{14}$$

最后网络 E_n, D_e 的损失函数式(12)与辨别器的损失函数的负数式(14)加权相加, 得到整体损失函数:

$$L^*_{E_n,D_e} = (1-\alpha)L_{E_n,D_e} + \alpha L_c \tag{15}$$

这样 E_n、D_e 网络的训练就和辨别器形成了对抗关系。

2.6　异常判断

在 2.3 节所示 LSTM 模型训练完成后, 即可通过模型进行人群异常行为判断。设原图像帧序列为$\boldsymbol{x}_i = \{x_1, x_2, \cdots, x_k\}$, 序列经编码后得到潜在向量序列$\boldsymbol{e}_i = \{e_1, e_2, \cdots, e_k\}$, 再将前 k–1 个潜在向量输入到 LSTM 模块得到预测的潜在向量序列$\boldsymbol{e}_i' = \{e_2', e_3', \cdots, e_k'\}$, 最后解码得到重构帧序列$\boldsymbol{x}_i'=\{x_2', x_3', \cdots, x_k'\}$。通过原帧序列与重构帧序列, 可定义异常分数:

$$S = \sum_{i=2}^{k} \|\boldsymbol{x}_i' - \boldsymbol{x}_i\| \tag{16}$$

此外, 利用辨别器输出得到原帧评价分和重构帧评价分, 分别为

$$S_x = \sum_{i=2}^{k} D(\boldsymbol{x}_i, \boldsymbol{e}_i) \tag{17}$$

$$S_{x'} = \sum_{i=2}^{k} D(\boldsymbol{x}_i', \boldsymbol{e}_i') \tag{18}$$

整体异常分数由(16)~(18)整合得到:

$$S_{\text{all}} = S + \alpha \cdot |S_x - \beta \cdot S_{x'}| \tag{19}$$

其中 α, β 是可调参数。

通过以上公式能够计算当前序列的最后一帧的异常分数。与 ST-VAE 相似, 采用 ST-AVAE 进行异常判断同样需要寻找一个最佳的阈值, 通过为异常分数设定阈值能够判断当前时间对于图像是否存在异常, 令模型的异常判断准确率达到最

高。即对于阈值 T, $S_{\text{all}}>T$ 时，当前帧判断为异常。

空时对抗变分自编码器对抗网络训练过程算法描述如下。

算法　对抗网络训练过程算法

1) 初始化编码器 E_n, 解码器 D_e, 辨别器 D

2) 迭代 N·$\mathcal{R}$: N=12 000

3) 采样 M 个图像样本 $(x_1,x_2,\cdots,x_m)$

4) 编码器生成 M 个编码 $(z_1',z_2',\cdots,z_m')$

$$\boldsymbol{z}_i' = E_n(\boldsymbol{x}_i)$$

5) 编码器重构误差：

$$L_R(\boldsymbol{x}) = \|x - D_e(E_n(\boldsymbol{x}))\|_1$$

6) 先验概率$P(z)$采样 M 个编码 $(z_1,z_2,\cdots,z_m)$

7) 解码器生成 M 个图像 $(x_1',x_2',\cdots,x_m')$：

$$\boldsymbol{x}_i' = D_e(z)$$

8) 解码器重构误差：

$$L_{R'}(\boldsymbol{z}) = \|\boldsymbol{z} - E_n(D_e(\boldsymbol{z}))\|_1$$

9) 正则化项：

$$L_c(\boldsymbol{x},\boldsymbol{z}) = L_R(\boldsymbol{x}) + L_{R'}(\boldsymbol{z})$$

10) 更新辨别器 D:

$$L_D = -\frac{1}{N}\sum_{i=1}^{N} D(x_i, E_n(x_i)) + \frac{1}{N}\sum_{i=1}^{N} D(D_e(z), z)$$

11) 更新编码器 E_n、D_e:

$$L_{E_n,D_e} = -(1-\alpha)L_D + \alpha L_c$$

空时对抗变分自编码器对抗网络训练过程算法中分别加入了编码器和解码器的重构误差，并将两式求和作为对其约束，即正则化项L_c，进一步降低了重构误差，提升了模型的的重构精度。

3　实验结果及分析

3.1　实验设置

3.1.1　数据集

为了验证本文模型的有效性，采用了 UMN 公开数据集[23] 和采集的逃散事件视频对本文方法和主流方法 ST-AE[24], ST-VAE[25] 进行了对比。UMN 数据集由 3 段不同场景下的人群异常事件模拟视频组成，记录了俯视视角下人群在视野中央漫步到爆炸式逃散的模拟异常事件过程。此外，本文采集了人群逃散行为的异常事件视频，该视频数据集中包含同一场景、两种不用视角下 10 人的爆炸式逃散过程。

3.1.2　数据预处理

为提取时序信息，需要将视频数据分散成若干个视频块，每个视频块由 n 帧连续图像组成，以视频块作为网络输入数据。由于视频数据较大，若按连续 n 帧，即步长为 1 的方式合成视频块，将导致数据量过大。此外，由于部分数据集连续两帧之间人群变化较不明显，为更好地检测异常事件的发生，需对视频进行抽帧处理，按一定步长对原视频数据进行采样。本文采用步长为 2 的方式由原视频采样 12 帧的单位视频块，以获得更好的算法性能。在异常检测过程中，模型以某时刻的前 11 帧作为输入，预测该时刻是否发生异常事件。根据上述视频块采样方式对数据集进行划分，处理后的各数据集构成如表 1 所示。

表 1　训练集测试集划分

Table 1　Training set and test set division

数据集	组成部分	序列数
UMN	训练集	3 101
	测试集正常帧	730
	测试集异常帧	787
采集数据集	训练集	3 457
	测试集正常帧	176
	测试集异常帧	200

3.1.3　参数设置

实验中 ST-AE, ST-VAE, ST-AVAE 的网络配置如表 2 所示。

表 2　网络配置

Table 2　Network configuration

编码器	图像	Conv1	Conv2	Conv3	Conv4	Conv5	Conv6	LReLU	LSTM
	128×128×3	64×64×6	32×32×9	16×16×12	8×8×24	4×4×48	256	64	64
解码器	LSTM	reshape	Deconv1	Deconv2	Deconv3	Deconv4	Deconv5	Deconv6	
	64	2×2×48	4×4×48	8×8×24	16×16×12	32×32×9	64×64×6	128×128×3	

ST-AE 模型中，编码器的输入维度为 128×128×3，经过 5 层 3×3 的卷积核，1 层 1×1 卷积核，输出维度为 64。LSTM 的输入维度即为 64，隐藏层神经元数为 32。解码器输入维度为 64，输出维度为 128×128×3。

ST-VAE 模型中，编码器的输入维度为 128×128×3，输出维度为 64。LSTM 的输入维度即为 64，隐藏层神经元数为 64，总共两层。解码器输

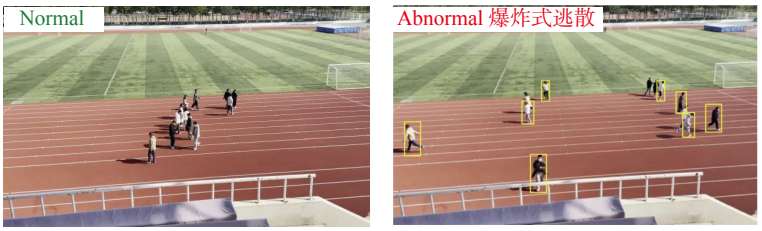
入维度为 64, 输出维度为 128×128×3。

ST-AVAE 模型中, 编码器的输入维度为 128×128×3, 输出维度为 64。LSTM 的输入维度为 64, 隐藏层神经元数为 64, 总共两层。解码器输入维度为 64, 输出维度为 128×128×3。辨别器输入维度为 (128×128×3,64), 输出维度为 1, 代表异常得分。

其中 VAE 模型学习率为 0.0002, LSTM 单元的学习率为 0.01, 学习回合数 epoch 定为 1, 批量数据 batch_size 为 64, 总批量 n_batch 为 12000。

3.2　实验结果

本文基于包含爆炸式逃散、同方向逃散两种异常行为的视频样本进行实验, 两种异常行为场景如图 7 所示。实验通过 ST-AVAE 模型重构误差随时间的变化验证方法框架的有效性, 并对原视频图像、人群密度图两种输入进行对比, 探究图像中不同因素对检测结果造成的影响。其中, 实验采用的人群密度图由原视频通过 DSNet[26] 模型生成。



(a) 采集视频数据下的人群正常到爆炸式逃散过程

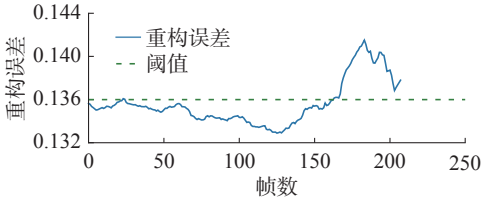
(b) 采集视频数据下的人群正常到同方向逃散过程

(c) UMN 数据下的人群正常到爆炸式逃散过程

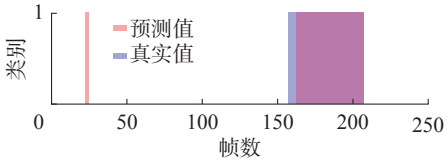
图 7　UMN 数据集与采集模拟异常视频数据包含的两种人群异常行为

Fig. 7　Two kinds of crowd abnormal behaviors included in UMN datasets and collected simulated abnormal video data

图 8 给出了将原视频图像输入模型时的重构误差变化情况与异常检测结果。图 9 给出了 UMN 公开数据集将原视频图像输入模型时的重构误差变化情况与异常检测结果。其中, 重构误差变化曲线中的横线代表模型得到最优准确率时对应的重构误差异常阈值; 异常检测结果示意图中, 模型对异常样本进行判断的预测值、真实值分别以红色、蓝色条带表示, 重合部分代表该时间样本预测正确。根据实验结果可知, 在爆炸式逃散和同方向逃散两种异常行为出现的时刻, 模型重构误差产生了明显地变化, 能够获得较好的预测效果。此外, 在爆炸式逃散初期出现了漏检的情况, 推测为人群四散开始时, 速度特征、密度变化特征均不明显, 导致出现漏检。在同方向逃散初期发现异常, 随后出现了少量的漏检情况, 推测为人群速度特征不显著, 被识别为人群正常移动。



(a) 爆炸式逃散的重构误差变化



(b) 爆炸式逃散的异常检测结果

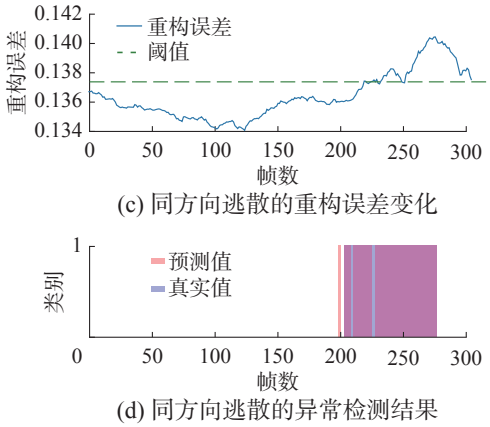


图 8　ST-AVAE 模型在采集数据集上重构误差变化及异常检测结果

Fig. 8　ST-AVAE model reconstructs error changes and anomaly detection results on the collected datasets

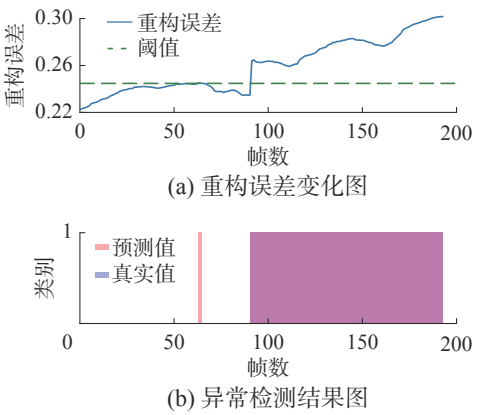


图 9　ST-AVAE 模型在 UMN 公开数据集上重构误差变化及异常检测结果

Fig. 9　ST-AVAE model reconstructs error changes and anomaly detection results on UMN public datasets

此外，为了进一步讨论图像的人工特征是否对基于重构的视频异常检测模型有所帮助，本文用密度特征图 DSNet[26] 替换原始图像作为输入，观察重构误差和异常检测结果，如图 10 所示。发现相较于原始方法，采用密度图的 ST-AVAE 的异常样本与正常样本的重构误差区分不够显著，预测结果准确度下降，出现了较多的漏检。说明颜色、外观、纹理、光影等信息为模型提供了更丰富的特征，保留了正常样本和异常样本的差异性，因此主要保留图像的密度特征对异常检测起负面作用。同时，根据异常预测结果示意图，模型在同方向逃散行为发生初期能够较准确地做出反应，但在一段时间后出现了漏报，推测为人群逃散方向较一致，造成画面被误检测为人群正常移动。

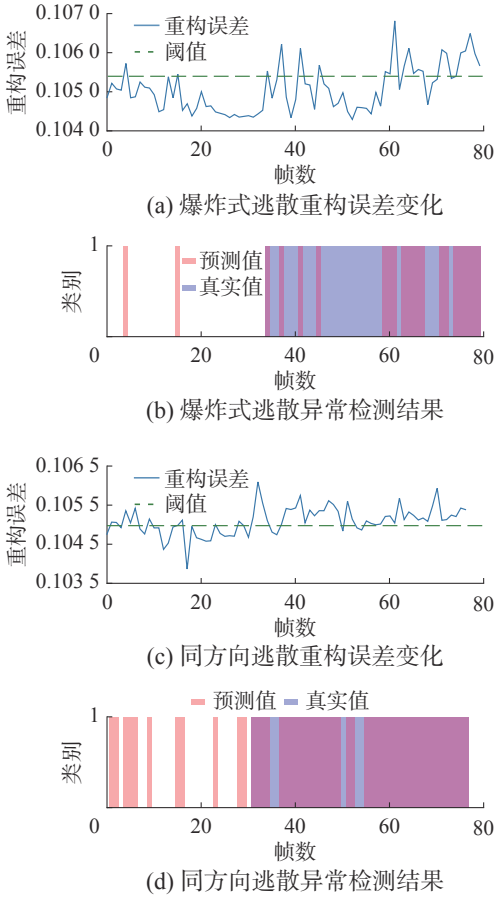


图 10　采集数据集在密度特征基础上重构误差变化及异常检测结果

Fig. 10　Reconstructs error changes and anomaly detection results of collected datasets based on artificial features

为进一步验证模型性能，本文选取各数据集上检测结果的精确率、召回率、F_1 值以及 AUC 值作为评价指标展开实验，对比 ST-AVAE 模型与作为其基础的 ST-AE、ST-VAE 模型的人群异常行为检测性能。

对比实验结果如表 3 所示，在采集数据集上，ST-AVAE 模型的召回率、精确率、准确率指标相较 ST-AE 模型在爆炸式逃散异常行为检测上分别提升了 11%、12% 以及 14%。在同方向逃散异常行为检测上分别提升了 2%、3% 以及 2%。相较 ST-VAE 模型在爆炸式逃散上分别提升了 4%、−3% 以及 1%，在同方向逃散上提升了 3%、2% 以及 2%；在 UMN 公开数据集上，ST-AVAE 模型相较 ST-AE 模型在爆炸式逃散异常行为检测上 3 种指标分别提升了 14%、19% 以及 16%，相较 ST-VAE 模型分别提升了−1%、10% 以及提升了 4%。通过实验结果可以发现，本文提出的 ST-AVAE 模型的召回率、精确率、准确率指标整体上相较其他方法有了明显提升，说明添加辨别器

模块, 采用对抗学习方法能够有效提升模型区分异常样本的能力。但是在融合了密度特征图的ST-AVAE 模型上的效果远不如在原图上的检测性能, 仅在同方向逃散上有良好表现, 推测为密度特征图受低分辨率影响, 不能很好地表示较为稀疏的人群, 由图 10(a) 所示, 在人群爆炸式逃散后出现了较多的漏报, 模型将异常行为识别为正常。

表 3　实验结果指标

Table 3　Experimental result indicators

场景	数据集	模型	F_1	AUC	召回率	精确率	准确率
爆炸式逃散	采集数据集	ST-AE	0.85	0.80	0.87	0.83	0.84
		ST-VAE	0.96	0.98	0.94	0.98	0.97
		ST-AVAE	0.99	0.97	0.98	0.95	0.98
	UMN数据集	ST-AE	0.80	0.80	0.83	0.81	0.83
		ST-VAE	0.94	0.97	0.98	0.90	0.95
		ST-AVAE	0.99	0.99	0.97	1	0.99
同方向逃散	采集数据集	ST-AE	0.96	0.97	0.96	0.96	0.96
		ST-VAE	0.96	0.97	0.95	0.97	0.96
		ST-AVAE	0.99	1	0.98	0.99	0.98

4　结束语

本文对基于深度学习的人群逃散异常行为检测方法进行了研究。针对现有方法未能充分解决样本不均衡带来的人群异常检测精准度低, 模型训练效率低等问题, 提出空时对抗变分自编码器的异常检测模型, 在 ST-VAE 模型基础上, 引入了 GAN 网络的辨别器结构, 并采用对抗学习方式提升模型对正常异常样本的分辨能力。通过与目前主流人群异常行为检测模型在公开数据集和采集数据的对比实验, 验证了对抗学习和空时信息帮助模型扩大了正常、异常样本重构误差差异, 提升了模型训练效率, 改善了一般基于重构的生成模型的过度泛化的问题。但该模型仍然存在对场景的依赖, 如何通过少量样本实现群体异常行为检测的域适应是未来的主要工作。

参考文献:

[1] LI Weixin, MAHADEVAN V, VASCONCELOS N. Anomaly detection and localization in crowded scenes[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(1): 18–32.

[2] XIE Shaoci, ZHANG Xiaohong, CAI Jing. Video crowd detection and abnormal behavior model detection based on machine learning method[J]. Neural computing and applications, 2019, 31(1): 175–184.

[3] HASAN M, CHOI J, NEUMANN J, et al. Learning temporal regularity in video sequences[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 733−742.

[4] CHONG Y S, TAY Y H. Abnormal event detection in videos using spatiotemporal autoencoder[C]//CONG F, LEUNG A, WEI Q. International Symposium on Neural Networks. Cham: Springer, 2017: 189−196.

[5] LUO Weixin, LIU Wen, GAO Shenghua. Remembering history with convolutional LSTM for anomaly detection[C]// 2017 IEEE International Conference on Multimedia and Expo. Hong Kong: IEEE, 2017: 439−444.

[6] 杨彪, 曹金梦, 张御宇, 等. 加权卷积自编码长短期记忆网络人群异常检测方法: CN108805015B[P]. 2021-09-03. YANG Biao, CAO Jinmeng, ZHANG Yuyu, et al. Weighted convolutional autoencoder-long short-term memory network-based crowd anomaly detection method: CN108805015B[P]. 2021-09-03.

[7] YAN Shiyang, SMITH J S, LU Wenjin, et al. Abnormal event detection from videos using a two-stream recurrent variational autoencoder[J]. IEEE transactions on cognitive and developmental systems, 2020, 12(1): 30–42.

[8] LIU Jie, SONG Kechen, FENG Mingzheng, et al. Semi-supervised anomaly detection with dual prototypes autoencoder for industrial surface inspection[J]. Optics and lasers in engineering, 2021, 136: 106324.

[9] LIU Wen, LUO Weixin, LIAN Dongze, et al. Future frame prediction for anomaly detection-A new baseline[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6536−6545.

[10] DONG Fei, ZHANG Yu, NIE Xiushan. Dual discriminator generative adversarial network for video anomaly detection[J]. IEEE access, 2020, 8: 88170–88176.

[11] NGUYEN T N, MEUNIER J. Anomaly detection in video sequence with appearance-motion correspondence[C]// 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2020: 1273−1283.

[12] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[EB/OL]. (2015-11-18)[2020-01-01]. https://arxiv.org/abs/1511.05644.

[13] 唐浩漾, 张小媛, 王燕, 等. 基于生成对抗网络的人体异常行为检测算法 [J]. 西安邮电大学学报, 2020, 25(3): 92–97.
TANG Haoyang, ZHANG Xiaoyan, WANG Yan, et al. Human abnormal behaviour detection algorithm based on generative adversarial nets[J]. Journal of Xi'an University of Posts and Telecommunications, 2020, 25(3): 92–97.

[14] LI Nanjun, CHANG Faliang, LIU Chunsheng. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes[J]. IEEE transactions on multimedia, 2021, 23: 203–215.

[15] 张冀, 曹艺, 王亚茹, 等. 融合 VAE 和 StackGAN 的零样本图像分类方法 [J]. 智能系统学报, 2022, 17(3): 593–601.
ZHANG Ji, CAO Yi, WANG Yaru, et al. Zero-shot image classification method combining VAE and StackGAN[J]. CAAI transactions on intelligent systems, 2022, 17(3): 593–601.

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

[17] MARKOVITZ A, SHARIR G, FRIEDMAN I, et al. Graph embedded pose clustering for anomaly detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10536−10544.

[18] GOYAL S, RAGHUNATHAN A, JAIN M, et al. DROCC: Deep robust one-class classification[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2020: 3711−3721.

[19] LI Chunliang, SOHN K, YOON J, et al. CutPaste: self-supervised learning for anomaly detection and localization[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 9659−9669.

[20] RUDOLPH M, WEHRBEIN T, ROSENHAHN B, et al. Fully convolutional cross-scale-flows for image-based defect detection[EB/OL]. (2021-10-06)[2022-12-01]. https://arxiv.org/abs/2110.02855.

[21] CARRARA F, AMATO G, BROMBIN L, et al. Combining GANs and AutoEncoders for efficient anomaly detection[C]//2020 25th International Conference on Pattern Recognition. Milan: IEEE, 2021: 3939−3946.

[22] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[23] UMN. University of minnesota dataset for detection of unusual crowd activity[EB/OL]. (2006-05-30)[2020-01-01]. http://mha.cs.umn.edu/proj_events.shtml#crowd.

[24] WANG Lin, ZHOU Fuqiang, LI Zuoxin, et al. Abnormal event detection in videos using hybrid spatio-temporal autoencoder[C]// 25th IEEE International Conference on Image Processing. Athens: IEEE, 2018: 2276−2280.

[25] AN J, CHO S. Variational autoencoder based anomaly detection using reconstruction probability[J]. Special lecture on IE, 2015, 2(1): 1–18.

[26] DAI Feng, LIU Hao, MA Yike, et al. Dense scale network for crowd counting[EB/OL]. (2019-06-24)[2020-01-01]. https://arxiv.org/abs/1906.09707.

作者简介：

邢天祎，硕士研究生，主要研究方向为模式识别与智能系统。

郭茂祖，教授，博士生导师，博士，中国人工智能学会机器学习专委会常委、中国建筑学会计算性设计学术委员会常委，主要研究方向为机器学习、智慧城市、计算生物学。北京建筑大学电气与信息工程学院院长，2019 年以第一完成人获吴文俊人工智能自然科学奖二等奖。发表学术论文 100 余篇。

赵玲玲，副教授，博士，中国计算机学会生物信息学专委会委员、中国建筑学会计算性设计专委会委员，主要研究方向为机器学习、城市计算、生物信息学。主持和参与国家自然科学基金青年基金、面上项目、重点项目 8 项。发表学术论文 40 余篇。