# Generative Neural Networks for Anomaly Detection in Crowded Scenes

Tian Wang, *Member, IEEE*, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, *Senior Member, IEEE*, Chang Choi, *Senior Member, IEEE*

**Abstract**

Security surveillance is critical to social harmony and people's peaceful life. It has a great impact on strengthening social stability and life safeguarding. Detecting anomaly timely, effectively and efficiently in video surveillance remains challenging. This paper proposes a new approach, called $S^2$-VAE, for anomaly detection from video data. The $S^2$-VAE consists of two proposed neural networks: a Stacked Fully Connected Variational AutoEncoder ($S_F$-VAE) and a Skip Convolutional VAE ($S_C$-VAE). The $S_F$-VAE is a shallow generative network to obtain a Gaussian mixture like model to fit the distribution of the actual data. The $S_C$-VAE, as a key component of $S^2$-VAE, is a deep generative network to take advantages of CNN, VAE and skip connections. Both $S_F$-VAE and $S_C$-VAE are efficient and effective generative networks and they can achieve better performance for detecting both local abnormal events and global abnormal events. The proposed $S^2$-VAE is evaluated using four public datasets. The experimental results show that the $S^2$-VAE outperforms the state-of-the-art algorithms. The code will be available publicly at https://github.com/tianwangbuaa/.

***Index Terms*—Spatio-temporal, anomaly detection, Variational AutoEncoder, loss function,**

## I. Introduction

VIDEO surveillance is a key tool to maintain the security and stability of public scene [1, 2]. Densely crowded environments (such as shopping centers, train stations, etc.), are equipped with CCTV cameras to meet the increasing challenges of security issues in these public areas. The surveillance systems generate a large amount of video data. Detecting abnormal events timely, effectively and efficiently from a large amount of video data, without human interaction and monitoring, has become a crucial task in video surveillance.

In video surveillance, abnormal events can be classified into *global abnormal event* (GAE) or *local abnormal event* (LAE) [3, 4]. We assume that abnormal events happen in the foreground. Most of current research focuses on detecting abnormal events from foreground. The detection of GAE is to identify the frames with an anomaly, while the task of detecting LAE, beyond identifying the frames with an anomaly, is to locate the individuals with abnormal behaviors in the frames. It is more challenging to detect LAE than GAE.

In this paper, we aim to improve the detection of the LAE and GAE. To this end, we propose to use a self-supervised learning method so that the detection task can be achieved more accurately and efficiently. The proposed algorithm, called $S^2$-VAE, includes 2 stages: the first stage is a shallow network, called $S_F$-VAE, with a low resolution input. And the second stage is a deep neural network, called $S_C$-VAE, with a high resolution input. The shallow network $S_F$-VAE was designed to filter out some palpable normal samples quickly, so that the next stage network $S_C$-VAE can learn a model from the remaining samples more effectively and more efficiently.

Inspired by the *Gaussian mixture model* (GMM), we design $S_F$-VAE, a new *Variational AutoEncoder* (VAE) model, so that the GMM-like distributions can be learned with $S_F$-VAE for the raw input data. In our experiments, this $S_F$-VAE is used to learn several latent variables to overcome the limitation of a single latent variable in traditional VAE. The purpose of using $S_F$-VAE is to filter out some obvious normal samples from the original samples, which can significantly reduce the training and testing time in the next stage.

In the second stage of $S^2$-VAE network, the remaining samples are firstly enlarged, and the enlarged samples are fed into $S_C$-VAE. This $S_C$-VAE, is a deep generative network with skip-connection between downsampling layers and upsampling layers. The convolutional operation in $S_C$-VAE can learn hierarchical features and a local relationship from the input, which can not be achieved by the fully connected layers in $S_F$-VAE. This deep $S_C$-VAE network can also integrate low/mid/high level features, and therefore it has stronger learning ability than shallow networks. Finally, from the information theory, the fusion of low-level and high-level information achieved by skip-connection can reduce the information loss caused by the transmission across layers in the generative network. From the feature representation perspective, the low-level feature can be treated as the auxiliary feature to the high-level feature [5, 6].

We show how the proposed $S^2$-VAE can be used for anomaly detection in video data in the experiments. Four public datasets are used to evaluate the algorithm's effectiveness and efficiency by comparing with state-of-the-art approaches. From the experimental results, we find that our $S^2$-VAE outperforms the state-of-the-art algorithms consistently.

Corresponding author: C. Choi (email: enduranceaura@gmail.com).

T. Wang and M. Qiao are with School of Automation Science and Electrical Engineering, Beihang University, China (email: wangtian@buaa.edu.cn (OR wangtian8704@gmail.com), meinaqiao@buaa.edu.cn). Z. Lin is with School of Computing, Ulster University, United Kingdom (email: z.lin@ulster.ac.uk). C. Li is with College of Electrical and Information Engineering, Lanzhou University of Technology, China (email: xjtulice@gmail.com). H. Snoussi is with Institute Charles Delaunay-LM2S-UMR STMR 6279 CNRS, University of Technology of Troyes, France (email: hichem.snoussi@utt.fr). Z. Liu is with Nanjing University of Aeronautics and University of Luxembourg (email: sduliuzhe@gmail.com). C. Choi is with Computer Engineering and IT research Institute, Chosun University, Rep. of Korea (email: enduranceaura@gmail.com)

The contributions of this paper are as follows:

- a shallow generative neural network built based on VAE, called $S_F$-VAE network is proposed. This network can help to reduce unnecessary normal samples, which helps to improve the speed of the anomaly detection.
- a deep generative network with more powerful learning ability, called $S_C$-VAE, is proposed to detect the abnormal event from video data. This $S_C$-VAE network has a skipped encoder-decoder structure, with a build-in VAE. The $S_C$-VAE makes full use of the advantages of both CNN and VAE. The network fuses the feature between the encoder layer and the decoder layer, which helps to reduce information loss due to the transmission across layers.
- the proposed approach was evaluated by using four public datasets. The results show that the proposed approach outperforms state-of-the-art algorithms.

The rest of the paper is organized as follows. Section II reviews the related work. Section III presents our $S^2$-VAE. The performance of $S^2$-VAE is evaluated in Section IV. This paper is concluded in Section V.

## II. Related Work

The state-of-the-art methods of the abnormal detection can be categorized into: 1) motion based models, and 2) spatio-temporal approaches combining motion with appearance information.

In the motion based models, the trajectory based method was used to detect motions [7, 8], since such representations can preserve the temporal structure of the abnormal events. The computational cost rose significantly due to occlusion in complex scenes. Thus, the no-tracking based methods were favored. The descriptors such as quantized optical flow [9], social model [10, 11], co-occurrence matrix based on frame intensity [12, 13], sptiao-temporal context representations [14, 15], etc had been proposed. For instance, an algorithm monitoring optical flow in a set of fixed local spatial positions was presented in paper [16]. The sum of squared differences was transformed into a probability distribution. The likelihood of observations respected to the probability distribution of the observations was calculated, and the likelihood falling below a preset threshold was detected as an alert. The sparse reconstruction cost (SRC) model was introduced in paper [17] over the multi-scale histogram of optical flow. Due to the insufficient performance of huge training samples in paper [17], the weighted orthogonal matching pursuit was adopted in [18] to improve the ability of the model for handling large samples. With suitable communication technology, the anomaly detection method can used for application [19, 20]. The main limitation of the motion based approach is that it cannot detect abnormal events with a sequence of similar normal actions, and it cannot distinguish among the appearance characteristics.

The spatio-temporal approaches combining motion with appearance [21] have been very successful in anomaly detection [22, 23]. These models provided a more comprehensive representations than the motion based method. In paper [24] the video was described by the nearby spatio-temporal interest points (STIPs), then Gaussian process regression (GPR) was adopted to cluster, learn, and infer the appearance and position relationship of the STIPs, finally the abnormal event was detected with competing performance while maintaining lower space-time complexity. The mixture of dynamic textures (MDT) was proposed in paper [25]. Moreover, a hierarchical mixture of dynamic textures (HMDT) was proposed for handling the high computational cost of paper [25] later. The events of low-probability were handled using discriminant saliency. The high hierarchical levels and long-range dynamics are important for event representation. Although several models have already been proposed, handcrafted features meet the challenge of universality. The efficient and effective abnormal event detection method consisting of a feature descriptor with a suitable pattern classification method remains an open problem.

The most recent research in this area is driven by deep neural networks [26, 27], with some significant achievements in abnormal event detection [28, 29]. The work in paper [30] used both normal and abnormal events to construct the training samples, and the spatio-temporal information had been taken into account in a convolution neural network in order to fuse the appearance and movement information in video frames. The work in paper [28] first proposed a fully-connected autoencoder with the handcrafted histograms of gradients (HOG) and histograms of optical flows (HOF) features as input. Then in consideration of feature representation, the video clips were used as input, in order to extract features automatically by the fully convolutional autoencoder. Despite the better performance that deep neural networks gain compared with handcrafted features, the robustness of the feature representation is still needed to be improved.

It is recognised that the deep neural network, especially the generative models (e.g, VAE) can yield better performance for abnormal event detection. We aim to design new generative models to extract more robust features, so that the LAE and GAE can be detected simultaneously by using the same architecture.

## III. Model Architecture

This section presents our approach for abnormal event detection from video sequences. Fig. 1 presents the workflow and visualization of our approach, including Fully Convolutional Neural network (FCN) [31] for foreground extraction and our proposed abnormal detection of $S_F$-VAE and $S_C$-VAE. The first row in Fig. 1 is our network, and the second row is the samples and results from the network.

Suppose we have $N+1$ video frames $\{X_i\}_{i=1}^{N+1}$, the first step in this model is to extract the foregrounds $\{G_i\}_{i=1}^{N+1}$ from this $N+1$ frames by using FCN. The FCN used in this paper is FCN-16s, which is built based on VGG-16 and pre-trained on Pascal VOC 2012 [32]. Two consecutive foregrounds $G_i$ and $G_{i+1}$ are used to calculate motion feature with the optical flow algorithm [33], which results in a set of $N$ motion images $\{O_i\}_{i=1}^{N}$ represented by the Munsell Color System [33]. Now, both $G_i$ and $O_i$ will be used as input to $S_F$-VAE. The $S_F$-VAE

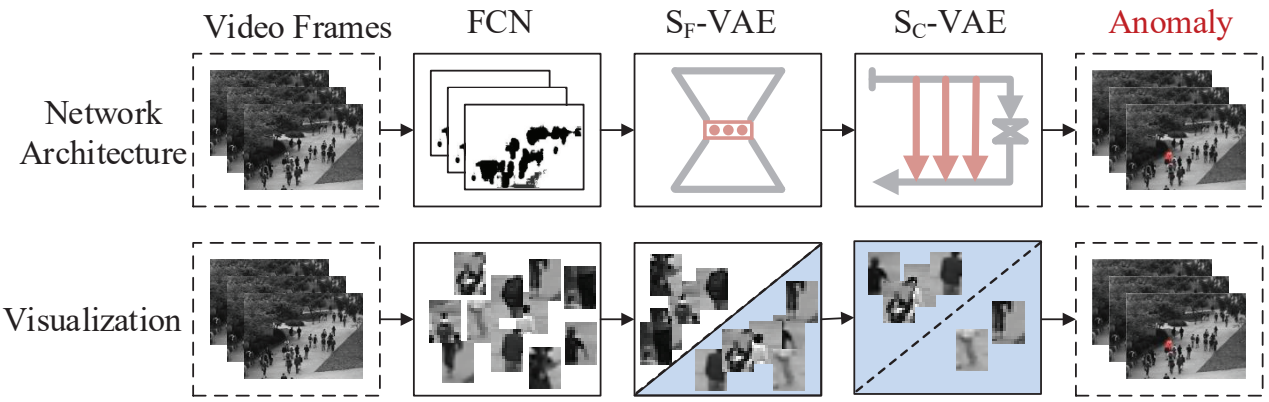



Fig. 1: The architecture for anomaly detection with a fully convolutional network (FCN), $S_F$-VAE and $S_C$-VAE. This architecture aims to detect LAE and GAE. For example, in LAE detection, it should be able to identify the abnormal objects.

will remove some unnecessary $G_i$ and $O_i$, and the filtered $G_i$ and $O_i$, as shown in blue in Fig. 1, will be used as input to $S_C$-VAE for detection.

In $S^2$-VAE, the $S_F$-VAE network, a shallow neural network, is designed to quickly filter some normal samples from the input sequences. The reduction of the training samples will not only decrease the training time of $S_C$-VAE, but also improve the robustness of the model. For the final stage, the $S_C$-VAE can extract abundant hierarchical features and allow the fusion of low-level and high-level features. It provides a more precise detection result.

### A. The $S_F$-VAE network

The proposed $S_F$-VAE network is used to learn a Gaussian mixture model. The study of VAE shows that a VAE is a perfect combination of neural network and variational inference [34]. From the neural network perspective, a VAE is an encoder-decoder architecture and from the variational inference perspective, it consists of an inference procedure and a generation procedure.

Let $x$ and $z$ be the inputs, where $x$ is the data input to VAE. $z$ as the latent representation of $x$, is learned by VAE. A VAE can be used to learn a Gaussian model such that $p(z|x) \sim N(\mu, \sigma^2 I)$ for approximating $x$. The loss function of VAE is shown as follows:

$$L = -E_{z\sim p_\theta(z|x)}[\log q_\phi(x|z)] + KL(p_\theta(z|x)||p(z)), \quad (1)$$

where $\theta$ and $\phi$ are the corresponding parameters to be trained in the encoder and decoder in the network. The first term is the reconstruction error between the input $x$ and the output decoded from $z$. The second term is the KL (Kullback-Leibler) divergence measuring the similarity between the distribution of $z$ and a known distribution where Gaussian distribution is mostly used.

Although VAE performs well in several applications, a VAE network with single latent variable may have limited capacity. Therefore, we propose to embed $n$ latent variables in a VAE network for abnormal event detection (shown in Fig. 2). The solid boxes represent all of the neurons in the corresponding layers, and the black dashed boxes represent the neurons of

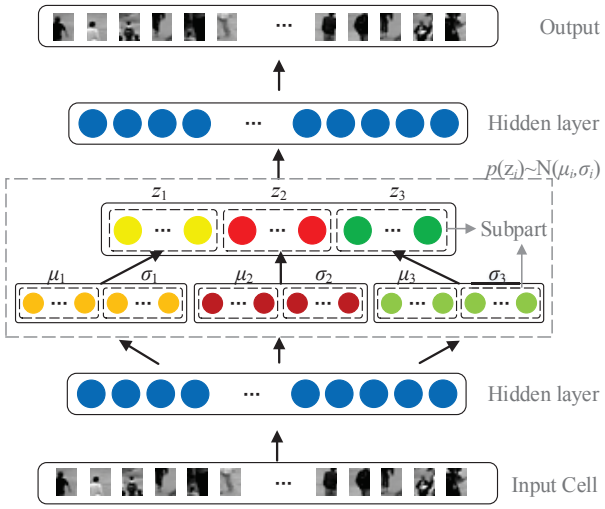



Fig. 2: The architecture of $S_F$-VAE in the first stage. The appearance of the region of the interest is taken as the training and testing samples in this figure. The motion based feature can also be handled in the same architecture.

each subpart. The large gray dashed box represents $n$ Gaussian components $p(z_i|x_i) \sim N(\mu_i, \sigma_i^2)$ where $1 \leq i \leq n$. We define the loss function of the $S_F$-VAE as:

$$\begin{aligned} L = \quad & -E_{z\sim p_\theta(z|x)}[\log q_\phi(x|z_1, \cdots, z_n)] \\ & + \tfrac{1}{n}\textstyle\sum_{i=1}^{n} KL(p_\theta(z_i|x)||p(z_i)), \end{aligned} \quad (2)$$

where the first term is the log-likelihood of the data, or the reconstruction error, and the second term is the average KL divergence between the distribution of the encoded $n$-latent variable and normal Gaussian distribution $p(z_i) \sim N(0, 1)$. Here, $\theta$ and $\phi$ are similar to the corresponding parameters in the Eq. 1.

The proposed $S_F$-VAE is inspired by the mixture of several Gaussian distributions. According to the theory of pattern recognition and machine learning, a simple Gaussian distribution does not have the ability to describe complex structures [35]. However, the mixture of Gaussian distribution is more powerful to fit the distribution of actual data. We

will demonstrate this proposed $S_F$-VAE's ability for modeling data in our experiments. The shallow network $S_F$-VAE was designed to filter out some palpable normal samples, so that the next stage network $S_C$-VAE can learn a model from the remaining samples more effectively and more efficiently.

### B. The $S_C$-VAE network

Despite the strength of Gaussian mixture like model in the first stage, it can only filter some normal samples out of the data samples. Since the $S_F$-VAE is still a shallow network, and the input is the direct flatten of the samples without considering the position relationship of the pixels. Then in the second stage of $S^2$-VAE, we build a deep network to extract more local relationship and hierarchical features from the input. And at the same time, in order to reduce the information loss across layers, we add a *skip connection* between low-level features and high-level features by using the concatenation of the feature map along the dimension of channel. Take the two feature maps shown in Fig. 3 with size $20 \times 16 \times 16$ as an example. The two feature maps are in the encoder step and decoder step. 20 and 16 are the height and width. The second 16 is the number of channels. After the skip connection which is labeled as 'M', the new feature map is with size $20 \times 16 \times 32$. It is the output of the skip connection and is also the input to the next decoder layer. The feature information is passed across the layers in the end-to-end network. The reason for adding the features from the encoder layer to the decoder layer is that information loss is inevitable in the decoding process. Therefore, it makes sense to combine the low-level features and the high-level features to reduce information loss.

The $S_C$-VAE network is built by combining U-net [36], and VAE (shown in the green rectangle in Fig. 3). The $S_C$-VAE network can not only extract local relationship and latent variables of the input data, but also integrate the feature maps with same resolution in the downsampling layers and upsampling layers, in order to obtain more accurate pixel-wise reconstruction.

Since the $S_C$-VAE network is to reconstruct the input data. The loss function for $S_C$-VAE for $N$ training samples is proposed as:

$$L=\frac{1}{N}\sum_{i=1}^{N}\left((x_i-\hat{x}_i)^2\right)+KL\left(p(z|x)||p(z)\right)+\gamma\left\|w\right\|_2^2, \qquad (3)$$

where the first term is the average reconstruction error of the training samples. $x$ is the input of the network, $x_i$ is the pixel value of one sample, $\hat{x}_i$ is the output of the network (the reconstruction of $x_i$). The second term limits the latent variable distribution to be a Gaussian distribution. The last term is a regularizer to avoid over-fit.

There are also other methods to reduce information loss including highway network [37], ResNet network [38] and so on. They are quite effective but they require very deep architecture. Our network is effective but it is not as deep as them [37], [38]. Therefore the skip connection proposed in this paper is more efficient for training our network. In addition, the built-in VAE network is not a general fully connected network consisting of layers with the same number of neurons, but is a reconstruction of its input. This is also beneficial to reduce information loss. On the other hand, the skip connection is an auxiliary feature added to the high-level features. The $S_C$-VAE network is a powerful generative network with less information loss and we will demonstrate its ability in the experiments.

### C. Anomaly detection

After we use $S_F$-VAE to process the input samples of $G_i$ and $O_i$, the output from the $S_F$-VAE network will be resized and the resized images will be the input to the $S_C$-VAE network. For example, if we have $16 \times 12$ images from $S_F$-VAE, we can resize them to $80 \times 60$, which is then fed to the $S_C$-VAE network. The convolution operation is similar to VGGNet [39]. Here, we have an example of how the $S_C$-VAE network operates on a resized images:
$I(80,60,3) \rightarrow CC(80,60,64) \rightarrow P(40,30,64) \rightarrow Z(40,32,64) \rightarrow CC(40,32,32) \rightarrow P(20,16,32) \rightarrow CC(20,16,16) \rightarrow P(10,8,16) \rightarrow CC(10,8,8) \rightarrow F(640) \rightarrow FC(6) \rightarrow FC(640) \rightarrow R(10,8,8) \rightarrow U(20,16,8) \rightarrow C(20,16,16) \rightarrow M(20,16,32) \rightarrow CC(20,16,32) \rightarrow U(40,32,32) \rightarrow C(40,32,32) \rightarrow M(40,32,64) \rightarrow CC(40,32,64) \rightarrow C(40,30,64) \rightarrow U(80,60,64) \rightarrow C(80,60,64) \rightarrow M(80,60,128) \rightarrow CC(80,60,128) \rightarrow C(80,60,3)$.

In this structure, $I(i,j,k)$ is the input data, meaning that $k$ channels of $i \times j$ pixels; $C$ is a convolution operation; $CC$ is to perform the same convolution operation twice. $P$ is max-Pooling; $Z$ is Zero-padding; $F$ is to flatten the feature map after the convolution operation, $FC$ represents fully-connected; $R$ is to reshape the output of the fully-connected layer to a suitable format as input to the latter operation; $U$ is Upsampling; $M$ is to concatenate additional link between the downsampling layers and upsampling layers, as shown in the red rectangle in Fig. 3. This operation concatenates the low-level features and high-level features which have the same resolution.

For accurate detection of an abnormal event, we use both motion and appearance features of the samples. In order to extract robust features, we train the network in every stage twice, one for motion feature extraction, and one for appearance feature extraction. The input samples are optical flow and intensity of the pixels, respectively. After getting the training samples, we then feed them into the $S^2$-VAE to represent both motion and appearance features. Since both of the $S_F$-VAE and $S_C$-VAE are generative models, the abnormal event is detected by the reconstruction error of the input with a threshold set by the highest reconstruction cost during training. The final decision is the union set of the motion and appearance anomaly detection results.

## IV. Experiments

In this section, we conduct experiments to validate the proposed networks. All the experiments are run on an NVIDIA GTX-1080 GPU. We use four benchmark datasets: UCSD [40], Avenue [15], UMN [41] and PETS [42].

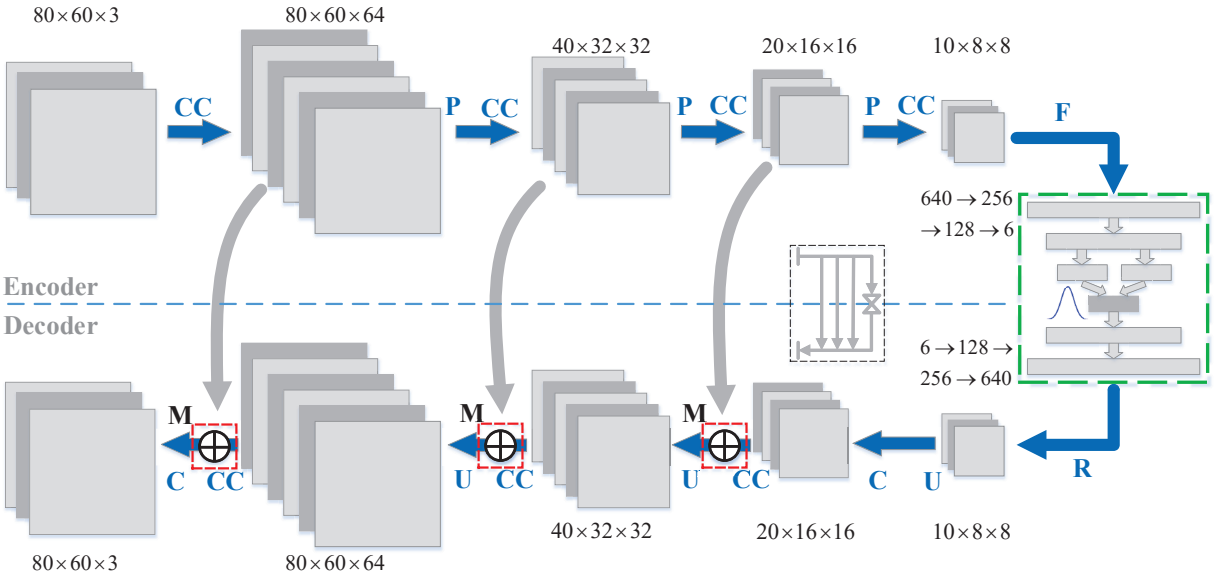



Fig. 3: The structure of the $S_C$-VAE network in the second stage. C: Convolution. CC: Convolution twice. P: max-Pooling. F: Flatten. R: Reshape. U: Upsampling. Z: Zero-padding. M: Merge link between the downsampling layers and upsampling layers.

### A. Pre-processing

For each video frame $X_i$, its foreground is extracted by using FCN as shown in Fig. 1. Fig. 4 shows an example of the foreground extraction for the original image.

For each foreground, blocks are extracted to cover every area in the foreground. To do so, we suppose each block has height $b_h$ and weight $b_w$ and each block contains cell units, where the size of each cell unit is $c_h \times c_w$. As such, one block will have $\frac{b_h}{c_h} \times \frac{b_w}{c_w}$ cell units. For example, in Fig. 4, the size of the foreground image is $158 \times 238$. If the size of each cell is $16 \times 12$, which is shown in the little red filled rectangle, then we can get at most $9 \times 19$ ($\frac{158}{16} \times \frac{238}{12}$) cell units in the block, which covers all the pixels between $(1, 1)$ and $(16\times9, 12\times19)$.

In order to cover the remaining pixels, we shift the block by a stride of 2 pixels to obtain different blocks so that all the pixels will be covered by a set of blocks. For example, in Fig. 4, the remaining pixels at the right and at the bottom are 10 pixels and 14 pixels respectively. We will shift the block to the right by 2 pixels to cover the pixels between $(1, 1+2)$ and $(16 \times 9, 12 \times 19 + 2)$, which results in a new block. By continuing this shift to the right or to the bottom, we could obtain 35 blocks in total, so that all the pixels will be covered at least by one block.

In our experiments, we only keep the cells whose area is covered at least $40\%$ with the foreground. And, the size of the cell is defined as $16 \times 12$ empirically so that it can cover an action and at the same time reduce the scale of the training samples. The extracted cells will then be used for LAE detection.

### B. Evaluation criteria

For abnormal event detection, the frame-level criterion is commonly used to evaluate both GAE and LAE detections.

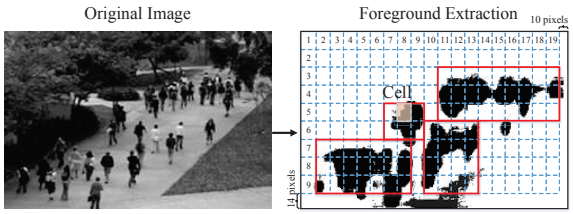



Fig. 4: The pre-processing for foreground detection and cell extraction.

But this is not good enough for LAE evaluation. The pixel-level criteria have been proposed for evaluating LAE detection [25].

**Frame-level Criterion:** The frame-level criterion is the accuracy of detecting abnormal events. A frame is classified as abnormal if an abnormal event is found in this frame, and the frame-level criterion only takes the whole frame into account. For frame-level evaluation, the *equal error rate* (EER), a trade-off between accuracy and recall, and the *Receiver Operating Characteristic* (ROC) curve, will be used. It is defined as the percentage of misclassified frames when the false positive rate equals the false negative rate.

This frame-level criterion is not an accurate evaluation method. For example, for an abnormal frame with a car on a walkway street as the abnormal event, a model may correctly classify this frame as abnormal but this decision was made based on the wrong detection that classifies a walking person as abnormal. Therefore, the frame-level is not accurate enough to locate a local abnormal event. As a result, we need pixel-level to fill this gap.

**Pixel-level Criterion:** This is to locate the abnormal events in a frame, rather than just tell if the frame contains abnormal events. In this case, a frame is classified as abnormal only if

the detected abnormal events have more than $40\%$ overlapping with the pixel-level ground truth. So the pixel-level criterion is a more accurate measure for evaluating the quality of the algorithm.

For pixel-level evaluation, the ROC curve, *Rate of Detection* (RD), and *Area Under receiver operating characteristic Curve* (AUC) are used. The RD is defined as the detection rate at equal error. The AUC is the area under the ROC curve. Therefore, if an algorithm is robust enough, then it will have low EER, high RD and high AUC.

### C. Experimental results

In this section, we will introduce the procedure and performance comparison of the experiment. The result estimation includes the comparison between the performance of networks with different architectures, and the comparison between proposed $S^2$-VAE network and state-of-the-art methods. In the $S_F$-VAE, 3 latent variables are constructed. For the network we propose, we make experiments on different networks similar with the proposed networks, and compare the results among them to make sure the proposed networks have gained the best performance in the aspect of the network. For comprehensive comparison, we compare our algorithm with autoencoder based model such as Conv-AE [28], and other state-of-the-art methods, such as Sparse [17], MDT [25], SF [43], MPPCA [44], MPPCA+SF [25], Adam [16], Feng [45] and so on. And this is to make sure the proposed algorithm has outperformed others in the detection of abnormal event in the aspect of algorithm.

*1) The UCSD dataset:* This is an LAE detection dataset, containing sequences taken on a walkway street by a stationary camera [40]. The density of the pedestrians varies from low to high. Each sequence contains 200 frames, and the resolution of each frame is $158 \times 238$. The normal events used for training are human walking, while the abnormal events are the frames with moving bikes, cars, wheelchairs and so on.

For the UCSD PED1 dataset, we first extract the foreground information by FCN network. Then we extract foreground blocks based on the cell units with size $16 \times 12$ in each frame and calculate their optical flow images. For each sequence in the dataset, as there are 200 frames, optical flow of 199 frames are extracted. We can get $9 \times 19 \times 199 = 34,029$ cells in the block of one position. After foreground extraction, we can get about 11,000 cells, including both normal and abnormal cells. Then for each cell unit, their raw pixels and their optical flow images are first fed to the $S_F$-VAE network to filter out some normal samples in the first stage, respectively. After $S_F$-VAE, we enlarge the height and weight of the remaining samples to $80 \times 60$, which are input into $S_C$-VAE in the second stage. The final decision is made based on the union of the motion feature and the appearance feature. The activations used in all of the neural network are Relu [46], and the optimizer is Adam with learning rate of $1e-4$. The results are shown in the 3D figures Fig. 5. In this dataset, each frame has 35 blocks, meaning that a pixel will be contained in at most 35 cells. The value for each pixel in Fig. 5 (e,f,g,h) is calculated based on the number of cells which contains the pixel and are classified as abnormal. It is obvious that pixels of the hikes in the 3D figures are the objects which are identified as abnormal.

TABLE I: The network comparison in the first stage. Comparison of our $S_F$-VAE network with general VAE networks.

| **Stage 1** | UCSD result | |
|---|---|---|
| Filter rate | $S_F$-VAE | VAE |
| | **5.7778** % | 1.1858 % |

TABLE II: The network comparison in the second stage. Comparison of our $S_C$-VAE network with other similar networks.

| **Stage 2** | UCSD result | | |
|---|---|---|---|
| Pixel-level AUC | $S_C$-VAE | No skip | FC |
| | **0.9425** | 0.7629 | 0.9303 |

To show the advantage of using both $S_F$-VAE and $S_C$-VAE networks, we do experiments to prove the effectiveness of the proposed networks. The results are shown in Table I and II. In Table I, we compare the performance on the proposed $S_F$-VAE with general VAE network by using the filter rate, where the filter rate is the proportion of filtered normal samples in all of the testing samples. We find that $S_F$-VAE has higher filter rate than the normal VAE. For stage 2, after using the $S_F$-VAE network for the first stage, we compare the performance on the proposed $S_C$-VAE with networks: 1) without the skip-connection in $S_C$-VAE; 2) without VAE, namely $F(640) \rightarrow FC(640) \rightarrow FC(640)$ in the architecture. The ROC of them is shown in Fig. 6 and the AUC is shown in Table II. The pixel-level AUC of the 3 networks is 0.9425, 0.7629, and 0.9303, respectively, which proves the advantage of using the $S_C$-VAE network.

We also compare the proposed approach with the state-of-the-art algorithms, shown in Fig. 7 for the ROC curve, in Table III for EER, RD and AUC. Our approach has lower EER, higher RD and higher AUC, compared to state-of-the-art

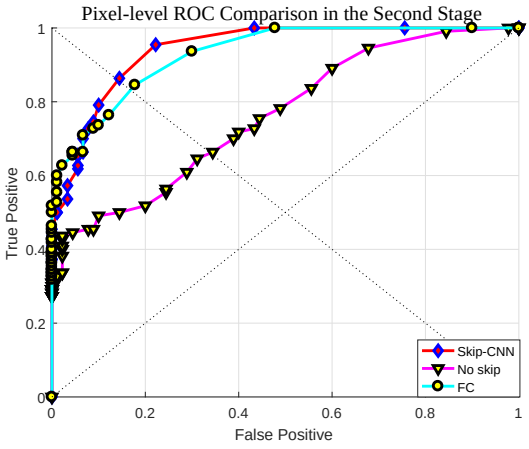



Fig. 6: The network comparison in the second stage on the UCSD dataset. Pixel-level ROC comparison between the $S_C$-VAE and other similar networks.

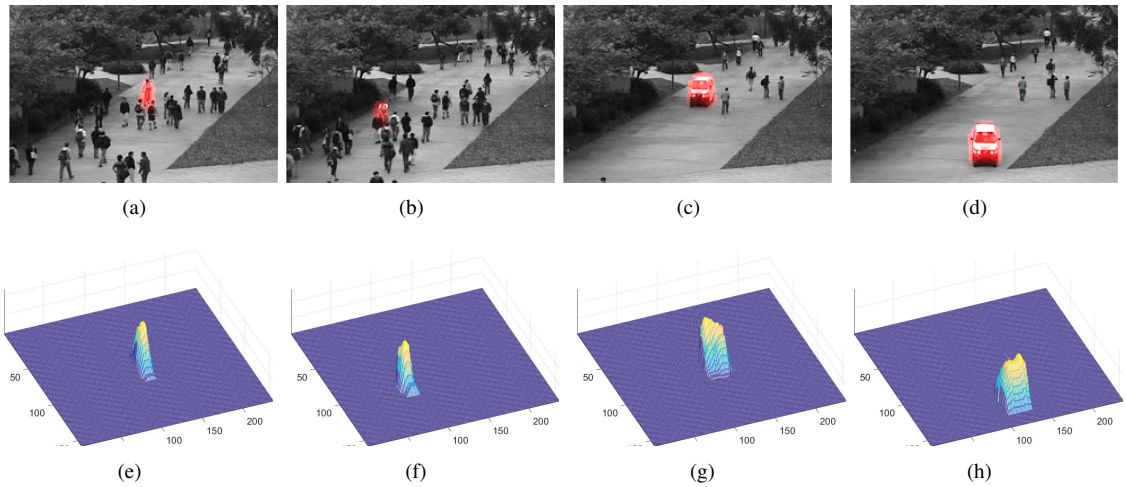


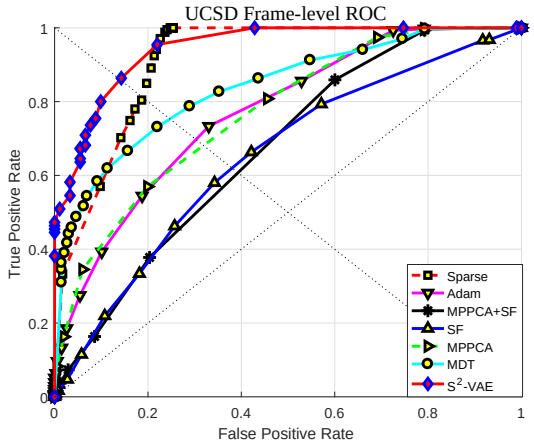

Fig. 5: Examples of LAE detection via our proposed algorithm. The hikes in (e), (f), (g) and (h) indicate objects being identified as abnormal. The values of the hikes are the number of times for a pixel being identified as abnormal in cells.

TABLE III: The algorithm comparison in our experiment. Comparison of our method with state-of-the-art methods for LAE of UCSD PED1 dataset. The best performances are shown in bold font. The *F* and *P* in brackets represents that the criterion is for the frame-level or the pixel-level.

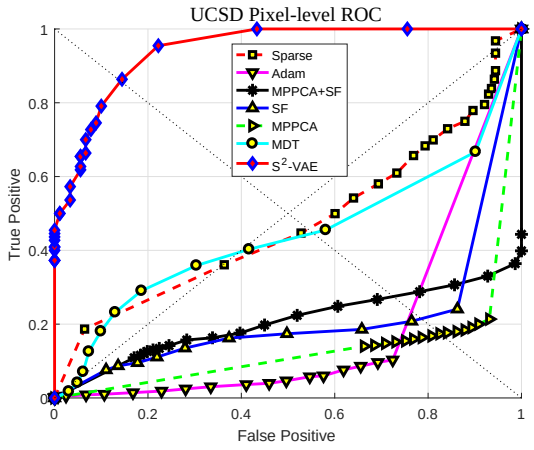
| Method | Evaluation Criteria | | |
|---|---|---|---|
| | EER (*F*) | RD (*P*) | AUC (*P*) |
| Sparse [17] | 19 % | 46 % | 46.1 % |
| Adam [16, 25] | 38 % | 24 % | 13.3 % |
| MPPCA+SF[25] | 32 % | 27 % | 21.3 % |
| SF [43] | 31 % | 21 % | 17.9 % |
| MPPCA [25, 44] | 40 % | 18 % | 20.5 % |
| MDT [25] | 25 % | 45 % | 44.1 % |
| HOG+HOS [12] | 27.02 % | 78.87 % | – |
| Conv-AE [28] | 27.9 % | – | 81.0 % |
| Lu [15] | – | 59.1% | 63.8% |
| sRNN [9] | 12.5 % | – | 89.9 % |
| Feng [45] | – | 64.9 % | 69.9 % |
| $S^2$-VAE (ours) | **14.3** % | **87.4** % | **94.25** % |

methods. In the $S^2$-VAE, on one hand, the $S_C$-VAE exploits robust feature extraction of CNN and data representation of VAE; on the other hand, the skip connections designed in the $S_C$-VAE can reduce information loss to gain a finer reconstruction of the input. Also, the first stage of $S_F$-VAE contributes to the detection of abnormal events by filtering out some normal samples effectively and reduce the number of input samples to the $S_C$-VAE. Thus, the performance of $S_C$-VAE can be improved by training without unnecessary normal samples.

*2) The Avenue dataset:* The Avenue dataset is an anomaly detection dataset provided by Lu et al. [15]. Since the ground truth of the Avenue dataset has been labeled by rectangles, it can be treated as an LAE dataset. There are 16 video clips for training, and 21 video clips for testing. The abnormal events




(a) Frame level ROC




(b) Pixel level ROC

Fig. 7: ROC comparisons of UCSD PED1 dataset. (a) Frame-level ROC for UCSD dataset. (b) Pixel-level ROC for UCSD dataset. The ROC of the compared algorithm is extracted from [17, 25].

TABLE IV: The algorithm comparison in our experiment. Comparison of our method with state-of-the-art methods for LAE of Avenue dataset. The best performance is shown in bold font.

| Method | AUC |
|---|---|
| Conv-AE [28] | 70.2 % |
| Lu [15] | 80.5 % |
| sRNN [9] | 81.71 % |
| Spatiotemporal-AE [47] | 80.3 % |
| Feng [45] | 75.4 % |
| $S^2$-VAE (ours) | **87.6** % |

TABLE V: The algorithm comparison in our experiment. Comparison of our method with state-of-the-art methods for GAE of UMN dataset. The best performances are shown in bold font.

| Method | AUC | | |
|---|---|---|---|
| | lawn | indoor | plaza |
| Social Force [43] | 96 % | | |
| NN [17] | 93 % | | |
| HOG+HOS [12] | 97.02 % | | |
| SRC [17] | 99.5 % | 97.5 % | 96.4 % |
| HOFO [9] | 98.45 % | 90.37 % | 98.15 % |
| CLP [48] | 98.72 % | 95.21 % | 99.34 % |
| $S^2$-VAE (ours) | **100** % | **99.92** % | **99.51** % |

include running, abnormal direction and so on. The resolution of each frame is $360 \times 640$. Compared with UCSD dataset, the Avenue dataset has higher resolution. The comparison, according to frame-level AUC, between our $S^2$-VAE with other algorithms is shown in Table IV. As can be found in the table, our method has better results than state-of-the-art methods.

*3) The UMN dataset:* This is a GAE detection dataset with three scenes: lawn, indoor and plaza. The image resolution of the dataset is $240 \times 320$. The global frame is handled by the proposed $S^2$-VAE method. In this dataset, the normal scenes are the events of people walking around, while abnormal scenes are the events of people running.

For the UMN dataset, as the behaviors (people running) of the abnormal events are similar in the scenes, we aim to train a model only on the lawn scene and then transfer this model to the indoor and plaza scenes. We show the experimental results in Table V. From the results, we find that our approach gains higher AUC, which also means our model is transferable.

*4) The PETS dataset:* This is a GAE detection dataset, captured by multiple cameras. The image resolution of PETS dataset is $576 \times 768$. This dataset has been applied to different tasks: event recognition, tracking, etc [42]. There are 2 different scenarios of abnormal events in this scene. In the first scenario, the normal events are defined as people walking in different directions, while the abnormal events are defined as people gathering and walking ahead in the same direction. In the second scenario, the normal events are defined as people walking in one queue, while the abnormal events are defined as people leaving the queue.

TABLE VI: The algorithm comparison in our experiment. Comparison of our method with state-of-the-art methods for GAE on the *Time14-17* scene and the *Time14-31* scene in the PETS dataset.

| Method | Detection accuracy | |
|---|---|---|
| | Time 1417 | Time 1431 |
| DT [49] | 93.8 % | |
| BoTG [49] | 91.2 % | |
| HOFO [9] | 97.8 % | 94.6 % |
| $S^2$-VAE (ours) | **99.3** % | **98.8** % |

For the PETS dataset, since the abnormal events are different in different scenarios, we train the model by the normal samples in each of the scenarios. Similar to UMN dataset, the proposed algorithm also works well on the PETS dataset. The results are shown in Table VI.

All of the experiments on different networks comparison demonstrate the superiority of our proposed network. The proposed network exploits the advantage of the robust feature extraction of CNN, the data representation of VAE, and the fusion of skip connections. This can reduce the information loss and gain a finer reconstruction of the input. As a result, the $S^2$-VAE gains excellent performance on abnormal event detection. And this superiority is also proved obviously by the latter experiments of both the comparison among similar networks and the comparison among state-of-the-art algorithms on the detection of LAE as well as GAE.

## V. Conclusion

Abnormal event detection from video sequences remains very challenging, due to the complexity of the video data. In this paper, a 2-stage algorithm, i.e. $S^2$-VAE, is proposed for the detection of both local abnormal event and global abnormal event. The proposed algorithm consists of 2 networks: $S_F$-VAE and $S_C$-VAE. The $S_F$-VAE network in the first stage is a shallow generative network for the powerful description of data distribution. It is used to filter out some unnecessary normal samples quickly. Then the $S_C$-VAE in the second stage is a deep generative network for accurately locating the abnormal events. The skip connection in $S_C$-VAE is to make the low-level features added to the high-level features as auxiliary features. In addition, the skip connection can also be viewed as the fusion of the information between the encoder and decoder, which can reduce the information loss across layers. And the VAE in the hidden layer also has the same effect. Finally, we show the effectiveness and efficiency of our proposed algorithm by the comparison on similar networks and the experiments on four public datasets.

## Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (61503017, U1435220,

61866022), the Aeronautical Science Foundation of China (2016ZC51022), the Fundamental Research Funds for the Central Universities (YWF-14-RSC-102), the SURECAP CPER project, the EU Horizon 2020 research and innovation programme under grant agreement (No. 690238) for DESIREE project, the UK EPSRC (No. EP/P031668/1), and the BT Ireland Innovation Centre (BTIIC) . and the Platform CAPSEC funded by Région Champagne-Ardenne and FEDER. Also,This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(Ministry of Science and ICT) (No. 2017R1E1A1A01077913).

## REFERENCES

[1] N. Y. Almudhahka, M. S. Nixon, and J. S. Hare, "Semantic face signatures: Recognizing and retrieving faces by verbal descriptions," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 706–716, 2018.

[2] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1359–1370, 2015.

[3] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[4] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 865–878, 2012.

[5] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing, "Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2008, pp. 69–82.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[7] T. Wang, M. Qiao, A. Zhu, Y. Niu, C. Li, and H. Snoussi, "Abnormal event detection via covariance matrix for optical flow based feature," *Multimedia Tools and Applications*, pp. 1–21, 2017.

[8] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2054–2060.

[9] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988–998, 2014.

[10] R. Chaker, Z. Al Aghbari, and I. N. Junejo, "Social network model for crowd anomaly detection and localization," *Pattern Recognition*, vol. 61, pp. 266–281, 2017.

[11] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino, "Abnormal crowd behavior detection by social force optimization," in *Human Behavior Understanding*. Springer, 2011, pp. 134–145.

[12] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments," *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2153–2166, 2015.

[13] Y. Shi, Y. Gao, and R. Wang, "Real-time abnormal event detection in complicated scenes," in *Proceedings of International Conference on Pattern Recognition (ICPR), Istanbul, Turkey*, 2010, pp. 3653–3656.

[14] H. Guo, X. Wu, S. Cai, N. Li, J. Cheng, and Y.-L. Chen, "Quaternion discrete cosine transformation signature analysis in crowd scenes for abnormal event detection," *Neurocomputing*, vol. 204, pp. 106–115, 2016.

[15] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013.

[16] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[17] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3449–3456.

[18] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, pp. 1851–1864, 2013.

[19] X. Su, H. Yu, W. Kim, C. Choi, and D. Choi, "Interference cancellation for non-orthogonal multiple access used in future wireless mobile networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 231, 2016.

[20] C. Choi, M. R. Ogiela, and H.-C. Chen, "Intelligent approaches for security technologies," *Concurrency and Computation: Practice and Experience*, vol. 30, no. 3, 2018.

[21] H. Shi, X. Li, K.-S. Hwang, W. Pan, and G. Xu, "Decoupled visual servoing with fuzzyq-learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 1, pp. 241–252, 2018.

[22] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[23] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, "Combining motion and appearance cues for anomaly detection," *Pattern Recognition*, vol. 51, pp. 443–452, 2016.

[24] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5288–5301, 2015.

[25] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern*

*Recognition (CVPR)*, 2010, pp. 1975–1981.

[26] T. Wang, Y. Chen, M. Zhang, J. Chen, and H. Snoussi, "Internal transfer learning for improving performance in human action recognition for small datasets," *IEEE Access*, vol. 5, pp. 17 627–17 633, 2017.

[27] T. Wang, Y. Chen, M. Qiao, and H. Snoussi, "A fast and robust convolutional neural network-based defect detection model in product quality control," *The International Journal of Advanced Manufacturing Technology*, vol. 94, no. 9-12, pp. 3465–3471, 2018.

[28] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 733–742.

[29] C. Jia, M. Shao, and Y. Fu, "Sparse canonical temporal alignment with deep tensor decomposition for action recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 738–750, 2017.

[30] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, 2016.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[32] VOC, "Pascal voc 2012 dataset. http://host.robots.ox.ac.uk/pascal/voc/voc2012/," 2012.

[33] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2432–2439.

[34] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[35] C. M. Bishop, *Pattern recognition and machine learning*. P101, springer, 2006.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[37] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] UCSD, "UCSD anomaly detection dataset, available from http://www.svcl.ucsd.edu/projects/ anomaly/dataset.html," 2010.

[41] UMN, "Unusual crowd activity dataset of University of Minnesota, http://mha.cs.umn.edu/movies/crowd-activity-all.avi," 2006.

[42] PETS, "Performance evaluation of tracking and surveillance (PETS) 2009 benchmark data. multisensor sequences containing different crowd activities. http://www.cvg.rdg.ac.uk/pets2009/a.html," 2009.

[43] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 935–942.

[44] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2921–2928.

[45] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, 2017.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 807–814.

[47] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.

[48] H. Fradi, B. Luvison, and Q. C. Pham, "Crowd behavior analysis using local mid-level visual descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 589–602, 2017.

[49] M. R. Khokher, A. Bouzerdoum, and S. L. Phung, "Crowd behavior recognition using dense trajectories," in *Proceedings of International Conference on Digital Image Computing: Techniques and Applications (DlCTA)*. IEEE, 2014, pp. 1–7.