Applied Intelligence (2023) 53:3240–3254
https://doi.org/10.1007/s10489-022-03613-1

Attention-based residual autoencoder for video anomaly detection

Viet-Tuan Le[1] · Yong-Guk Kim[1]

Accepted: 9 April 2022 / Published online: 25 May 2022

Abstract

Automatic anomaly detection is a crucial task in video surveillance system intensively used for public safety and others. The present system adopts a spatial branch and a temporal branch in a unified network that exploits both spatial and temporal information effectively. The network has a residual autoencoder architecture, consisting of a deep convolutional neural network-based encoder and a multi-stage channel attention-based decoder, trained in an unsupervised manner. The temporal shift method is used for exploiting the temporal feature, whereas the contextual dependency is extracted by channel attention modules. System performance is evaluated using three standard benchmark datasets. Result suggests that our network outperforms the state-of-the-art methods, achieving 97.4% for UCSD Ped2, 86.7% for CUHK Avenue, and 73.6% for ShanghaiTech dataset in term of Area Under Curve, respectively.

Keywords Anomaly detection · Residual autoencoder · Channel attention · Temporal shift · Anomaly video datasets · Unsupervised learning

1 Introduction

Anomaly detection in video surveillance is a popular research area in computer vision because of its diverse applications, such as traffic accident detection, criminal activity detection, or detecting illegal activities. And yet, detecting an abnormal activity among the vast normal situations is challenging. The first challenge is to collect and label all types of abnormal events since the frequency of normal events dominates that of abnormal events and often the abnormal parts are rare. The second challenge is their uncertain characteristic of abnormal events. For instance, an activity is regarded as anomalous in one context, but it can be a normal activity in the another case. When a pedestrian crosses the street in a crosswalk, the event is considered as a normal activity. However, the same activity is considered as abnormal when there is no crosswalk. Moreover, given that it is time-consuming and inefficient to watch and analyse the massive amounts of surveillance videos by human, an automatic anomaly detection system is essential for analysing and detecting abnormal events in surveillance videos.

As the goal of frame-level anomaly detection in videos is to identify the frames that contain different spatial and motion information, an anomaly detection model that has been trained using only normal samples (or frames) to learn a generic distribution of normal events cannot represent unseen events or activities which are considered as anomalies. However, abnormal frames can be distinguished using the reconstruction/prediction error between the ground truth sample and the reconstructed/predicted output while testing.

In video anomaly detection, motion information is one of the most important criteria by which one can make a decision whether it is normal or abnormal. Some existing approaches use a two-stream network [7, 14, 15] for anomaly detection, including a spatial stream and a temporal stream. The former learns the spatial structure of input frames while the latter leverages the optical flow between neighboring frames. However, the extraction of optical flow costs an extra computational power. Another approach uses a recurrent neural network, such as a variational LSTM [16, 22, 38] to model temporal motion information, although the model becomes too complex as the number of the stacked layer is increased [33, 35].

✉ Yong-Guk Kim
ykim@sejong.ac.kr

Viet-Tuan Le
tuanlv@sju.ac.kr

[1] Department of Computer Engineering, Sejong University, Seoul, Korea

In machine learning, attention is a technique that imitates human cognitive attention, enhancing a part of input, such as an object, but neglecting the remaining parts. In the anomaly detection area, [39] showed that attention selectively applied to the foreground area, wherein dynamic objects were moving, enhanced the performance while neglecting the static background area.

To handle these issues, we introduce an Attention-based Spatio-Temporal NETwork (ASTNet)[1], which has an architecture of autoencoder for the efficient anomaly detection task. The proposed network aims to exploit both spatial and temporal features efficiently within a unified manner. So that, the extracted features by a Deep Convolutional Neural Network (DCNN) are fed into two parallel branches to exploit both spatial structures and motion features. Then, the spatio-temporal features are again fed into a decoder to predict the future frame. Contrast to the figure-ground separated application of attention [39], we propose a cascade attention model where a channel attention module is inserted at each layer of the decoder to better exploit the channel relationship of the features. The main contributions of our work can be summarized as follows:

- We propose an attention-based residual autoencoder for video anomaly detection, which encodes both spatial and temporal information in a unified way.
- The temporal shift is applied to model temporal information, since it provides high performance with a low computational cost.
- The channel attention is applied to exploit channel dependency in a cascade type within the decoder to predict the future frame more efficiently.
- Our model outperforms state-of-the-art performance on three standard benchmark datasets, even without using any optical flow detector.

The rest of this paper is organized as follows. An overview of related work is discussed in Section 2. Section 3 describes our proposed method. Detailed experiment results and discussions are given in Section 4. Finally, Section 5 concludes this paper.

2 Related work

Recently, anomaly detection has been attracted a lot of attentions of the researchers. There are roughly two representative approaches in the video anomaly detection: the reconstruction-based method and the prediction-based method.

Reconstruction-based method. With this, a model is trained to reconstruct the input frame. The most popular model among many is the autoencoder architecture, consisting of an encoder and a decoder: the former compresses the input into a lower-dimensional feature representation, and the latter reconstructs the output from the compressed representation as close to the input frame as possible. Then, the reconstruction error is used to distinguish the abnormal event from the normal ones since the normal events have the smaller errors, whereas the abnormal event has the bigger one.

To extract the appearance feature as well as the motion feature from the video input, some approaches [28, 30] learnt the normal events by using an autoencoder architecture, which utilised both the stacked convolutional neural network layers to learn the spatial structure and a stacked convolutional LSTM to learn the temporal representation. In some case [28], a human observer was used for validation as a continuous learning. Recently continual learning has been applied for video anomaly detection to deal with the forgetting problem happening while training deep neural networks. Doshi and Yilmaz [6] use a deep learning model to extract feature embedding for input video frames. A set of nominal feature vectors is stored in a memory module using the k-nearest-neighbors. This process is trained in multiple session for continual learning.

A two-stream model [14] was often used to capture both the appearance and motion information. Such a model typically had an architecture that included an autoencoder and a discriminator. The anomaly scores of the two streams were combined for more accurate decision. Similarly, Li et al. [15] introduced a two-stream network to encode the appearance and motion of normal events in videos. Each stream of the network included two spatio-temporal autoencoders using 3D video cuboids as input. The 3D video cuboids were stacked from multiple patches which were partitioned at the same location in continuous frames. To overcome the high computational cost of optical flow, Chang et al. [2] used two autoencoders to separately exploit spatial and temporal information of videos. The spatial autoencoder encoded the scenes and objects while the temporal one captured the movement information of the objects. Fang et al. [8] proposed a multi-encoder single-decoder model to encode both motion and content cues. The network had a motion encoder and two content encoders. The outputs of these encoders were concatenated and reconstructed by a decoder.

A 3D convolutional neural network had the capability for learning both spatial and temporal information corresponding to appearance and movement, respectively, in videos. Deepak et al. [4] showed that an encoder with a convolutional LSTM layer processed spatial information whereas a decoder captured temporal one. Recently, a deep autoencoder had been used to reconstruct the input. For instance,

[1] https://vt-le.github.io/astnet/

[1] introduced a probabilistic model using an autoregressive process to estimate the density in the latent vector, that was extracted by an encoder. In addition, [10] reconstructed the input using an autoencoder with a memory module. The memory contents were learnt during the training phase, and the model reconstructed a testing input using the memory, which was learnt from the normal samples. As a result, an abnormal event produced a large reconstruction error. On the other hand, [13] proposed a three-stage method, which required the less computational cost. The authors substituted the autoencoder with a single-hidden-layer feedforward neural network, that reconstructed the input frames by minimizing the reconstruction error with a less computation time.

The sparse coding-based anomaly detection approaches [25, 38] were to detect anomalies using a learnt event dictionary. In such a case, the normal events were reconstructed from a learnt dictionary with a small reconstruction error, while the abnormal event would lead to a large reconstruction error. Within this context, [25] proposed a sparse coding based deep neural network using the stacked recurrent neural networks to optimize the sparse coefficients, while [38] introduced an optimization network based on a novel LSTM network. A fast sparse coding network [32] adopted a two-stream neural network to extract the spatio-temporal features as it was a lightweight network to learn a normal event dictionary.

Prediction-based method. This approach utilises a few previous frames in predicting whether the future frame would be normal or abnormal. The basic assumption is that the normal event is predictable whereas the abnormal one is unpredictable [20]. The frame prediction approaches usually exploit both appearance and movement information of the given video since the input contains several consecutive frames, which include motion features.

Generative Adversarial Network (GAN), consisting of a generator and a discriminator, is one of the most popular network recently, and it can be used to generate the next frame for the video anomaly detection task. For instance, [20] used the U-Net as the generator in predicting the next frame and a patch discriminator was adopted to distinguish the frames generated by the generator. Zhou et al. [39] used the similar network architecture, wherein U-Net was employed as a generator and a patch network as a discriminator, was used to predict the future frame. Moreover, an attention-driven loss was used to deal with the imbalance problem between the foreground object and the static background typically appeared in the anomaly detection videos. Similarly, [36] integrated the segmentation map into the PSNR (Peak Signal to Noise Ratio) to assign different weights to the background and the foreground. They also proposed the patch-level loss in their prediction model to improve the quality of the foreground object. In addition, [16] used a generative model to predict the future frame. In this case, however, the original U-Net of the generator was replaced by a spatio-temporal U-Net, which was added three ConvLSTM layers in the middle of the U-Net to model temporal information. Lu et al. [22] combined variational autoencoder and ConvLSTM to predict the future frame. The ConvLSTM was used to represent the recurrent relationship among frames in the given video. Doshi and Yasin [5] predicted whether the future frame would be normal or abnormal using a GAN. In this case, an object detection system had been used to extract the location and the appearance feature. The reconstruction errors and extracted information of objects were computed using a statistical module to detect the anomalies.

Hybrid method. Tang et al. [29] combined a future frame prediction approach with a reconstruction approach to exploit advantages of the above mentioned methods. Two blocks of U-Net were connected in series: the first block was for predicting whether the future frame was normal or abnormal and the second for reconstructing the frame. On the other hand, [27] used dynamic skeleton features for video anomaly detection. The skeletal movements were decomposed into global body movement and local body posture, and then fed into two recurrent encoder-decoder network branches that were employed to reconstruct their own input and predict the future frame. Chang et al. [3] adopted a two-stream network that exploited spatial and temporal information. In the first stream, an autoencoder encoded spatial information while a motion autoencoder predicted RGB difference between the first and the last frame to obtain motion information in the second stream instead of computing optical flow with an expensive computation. On the other hand, object based multi-task learning [9] jointed three self-supervised and one knowledge distillation for anomaly detection in video. In each frame, object detection was carried out with a pre-trained detector. A sequence of detected objects from consecutive frames was fed into a 3D CNN and four 2D prediction heads to detect anomalous events.

Although two-stream network and 3D CNN has proved the capability to model motion information without computing optical flow, such improvement comes with the high computational cost. In this study, we propose a simple autoencoder architecture which includes an encoder to extract feature from input video frames and a decoder to generate the future frame in an unsupervised fashion since the training videos contain only the normal events. The temporal information is exploited by an effective temporal shift method which is inserted into the network at zero computation and zero parameters [19]. In [39], an attention map is learned to force the model focus on the foreground rather than the background. However, the attention is effective with single scene dataset, such as UCSD Ped 2, CUHK Avenue. To tackle the multi scene dataset problem such as ShanghaiTech, we propose a channel attention-based decoder which focus on important objects automatically while predicting the future frame.

3 Method

In this section, we present our framework for video anomaly detection in detail. As mentioned before, abnormal events are very rare in real-world scenarios. Therefore, it is difficult to collect and label training data that cover all types of anomalies. To deal with this problem, we propose an unsupervised learning method for detecting abnormal events in video.

2D CNN [1, 10] has been used for diverse video anomaly detection tasks and yet it cannot represent the temporal features very well. To handle this problem, some approaches [28, 30] combine a 2D CNN and a temporally recurrent network such as convolutional LSTM. Such a combination aims to propagate temporal information across frames. Nevertheless, the more layers the model has, the more complex the model is. Another type of method that tries to capture both spatial and temporal information from videos would be 3D CNN [4] with which both spatial and temporal features can be learnt although it takes lots of effort to train the network. A few recent state-of-the-art methods [2, 14, 15] adopt a two-stream neural network, which consists of a spatial stream and a temporal stream. The spatial stream exploits the appearance features while the flow stream captures the motion information, and yet the computation of optical flow is rather expensive.

Problem statement. We propose a network for video anomaly detection using the future frame prediction approach. The input of the network is a sequence of frames in a video and the network tries to predict the future frame [20]. Given several consecutive frames $I = \{I_1, I_2, ..., I_t\}$, the predicted frame is $\hat{I}_{t+1}$ and the ground truth frame of the predicted one is I_{t+1}. Then, the anomaly score can be calculated using the difference between the predicted frame $\hat{I}_{t+1}$ and the ground truth one I_{t+1}.

3.1 Network architecture

The overall structure of the proposed model is shown in Fig. 1 and it has the autoencoder architecture, consisting of an encoder and a decoder. The former is for capturing both appearance and motion information of the input video frames, and the latter is for predicting the future frame using the extracted spatio-temporal features with the encoder.

Encoder. From a given sequence of t frames, the high-level features can be extracted by using a deep and wide convolutional neural network, i.e. WiderResnet [34]. In order to exploit both spatial and temporal information of video frames, the last feature map obtained from the deep convolutional neural network is then passed through two branches, as illustrated in Fig. 1. In the temporal branch, temporal shift is applied to model temporal features over several input frames (Section 3.2), while the extracted features of input frames are concatenated to maintain the spatial information in the spatial branch (Section 3.3). Then, the outputs of two branches are combined using an element-wise sum and fed into the decoder to predict the corresponding future frame.

Decoder. The output of the encoder is then used as input of the decoder. The combined features are passed through the decoder to restore the details and the spatial resolution of the predicted frame. Each layer of the decoder is a sequence of blocks, including deconvolution, batch normalization, and Rectified Linear Unit (ReLU) activation function. To exploit the channel relationship of features, the channel attention is applied after each deconvolution block, described in Section 3.4. In addition, the output features of the channel attention are concatenated with the corresponding low-level features extracted by the deep convolutional neural network that have the same spatial resolution. The combined features are used in the next step. Then, they are deconvolved to upsample the features back to the input frame resolution.

3.2 Temporal branch

The temporal shift process [19] has been used in the video understanding area. In the present work, we would like to utilize the temporal shifting technique to exploit temporal information in the video anomaly detection task. The shift operation is performed along the temporal dimension. Some part of the channels is shifted to the next frame while keeping the remaining part, as illustrated in Fig. 2. Then, the feature of the current frame is combined with the feature of the previous one. For the given input feature maps $\mathbf{F}_{\mathbf{tem}} \in \mathbb{R}^{N \times T \times C \times H \times W}$, the output features are computed as:

$$\mathbf{F}'_{\mathbf{tem}} = Shift(\mathbf{F}_{\mathbf{tem}}), \tag{1}$$

where $Shift$ refers to the shift operation. In Fig. 2, input features consist of four frames $T = \{t_1, t_2, t_3, t_4\}$. Part of the channels of the current frame is shifted to the next frame. Note that part of the channels of frame t_2 is replaced by the part of a channel of frame t_1.

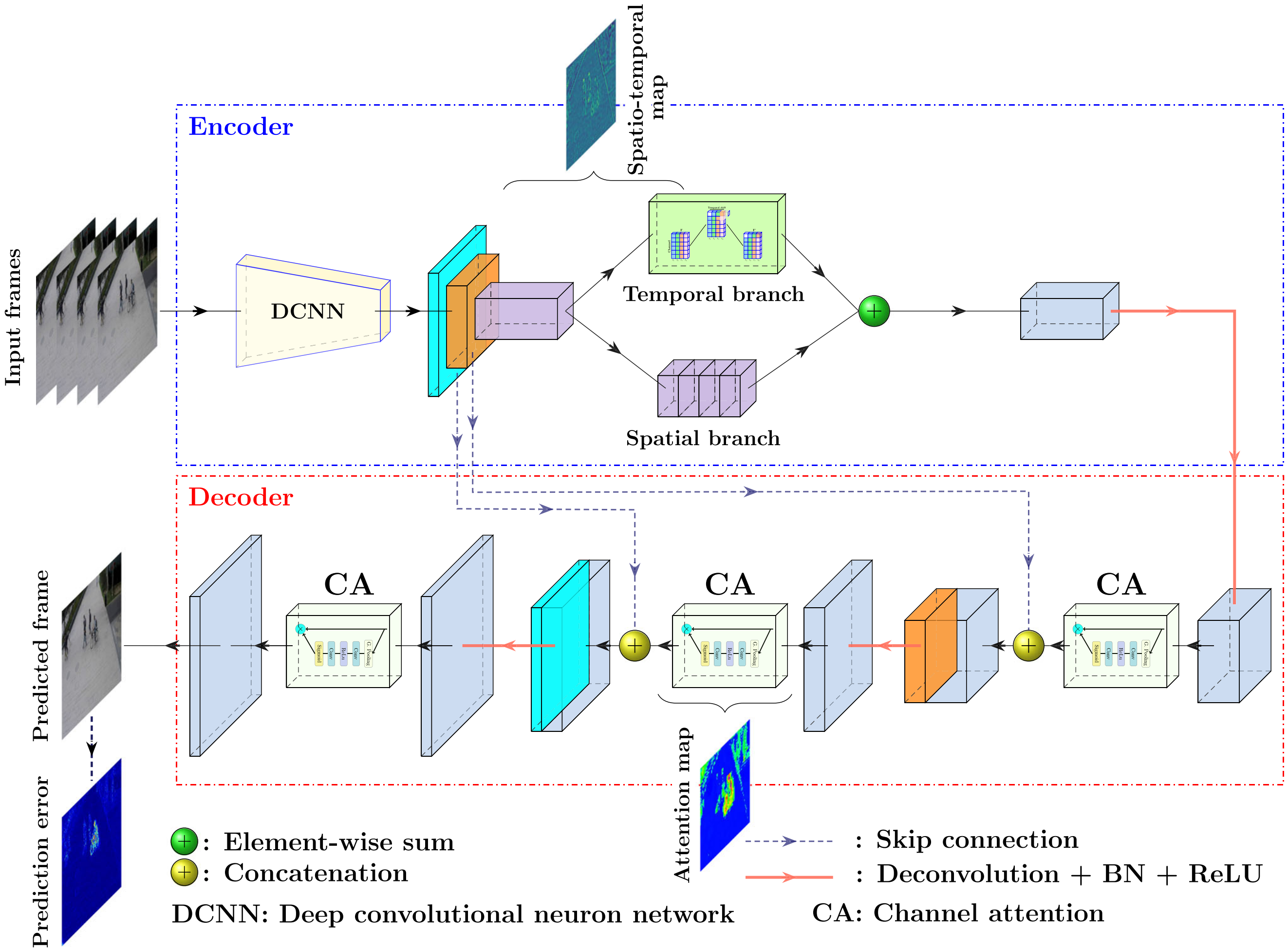
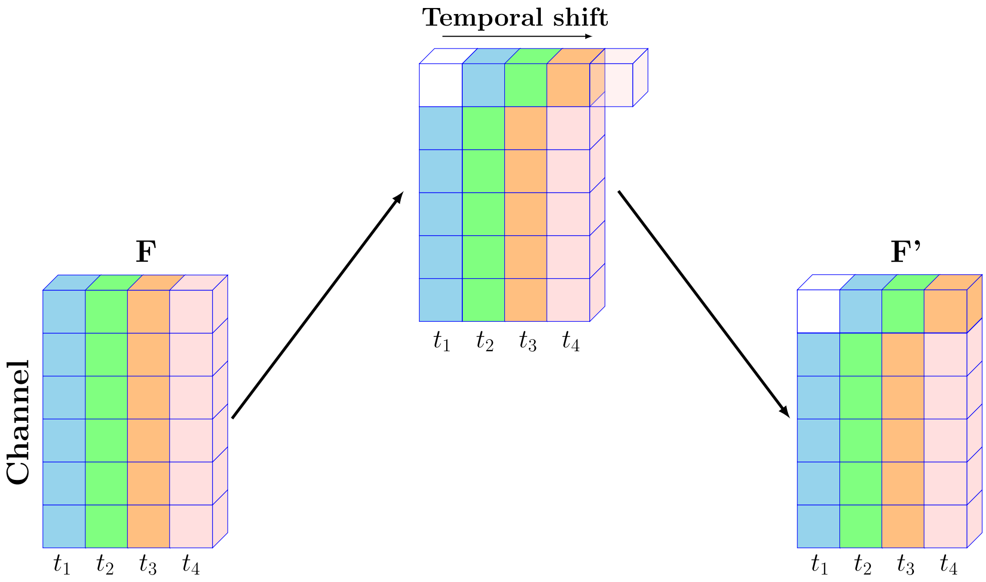


Fig. 1 The overall architecture of our network for video anomaly detection. Initially, a sequence of input video frames is fed into a DCNN to extract features. Then, the extracted visual features are passed through two branches to exploit further spatial information as well as temporal one respectively. The spatial and temporal features are combined and passed through three deconvolutional layers to generate a future video frame. Note that a Channel Attention (CA) is applied at each deconvolutional layer to exploit the channel dependency of the features in a cascade type to enhance the network performance

Fig. 2 The temporal shift. Given the feature map **F**, the output feature **F**′ is obtained by applying a temporal shift to exploit the temporal information. As illustrated, the features of different frames are described as different colors in each column. Part of the channel of frame t_1 (blue) is shifted to the next frame t_2 (green)



3.3 Spatial branch

In the spatial branch, the extracted features obtained from the deep convolutional neural network are aggregated across frames. To reduce computation complexity, we apply a 1 × 1 convolution on the combined features to reduce the number of channels since the aggregated features contain a large number of channels.

The features of the temporal and spatial branches are combined as follows:

$$\mathbf{F} = \mathbf{F_{tem}} + \mathbf{F_{spa}} \tag{2}$$

where $\mathbf{F_{tem}}$ and $\mathbf{F_{spa}}$ denote the output features of the temporal and spatial branches respectively.

3.4 Channel attention

In order to exploit channel dependency of the feature, channel attention [12, 31, 37] has been used in many fields. For instance, 'Squeeze-and-Exciation' [12] adopts global average pooling while CBAM [31] takes average-pooling and max-pooling in obtaining the channel-wise statistics. In our channel attention module, two convolutional layers are chosen like [37] instead of two fully-connected layers [12, 31].

After each deconvolutional layer, we apply channel attention for the feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$. The output feature $\mathbf{F}'$ is computed as follows:

$$\mathbf{F}' = \mathbf{F} \otimes s(\mathbf{F}), \tag{3}$$

where $s(\mathbf{F})$ refers to the channel attention, and ⊗ denotes element-wise product.

Channel Attention. The output of each deconvolutional layer is given as an input feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ of the channel attention module. In order to exploit channel dependency, a global average pooling is applied to the feature $\mathbf{F}$ [12, 37]. The output of the global average pooling is a vector $\mathbf{v}$ having C values. Then, a 1 × 1 convolution is applied to reduce the dimension with a reduction ratio r, followed by a rectified linear unit (ReLU) activation function δ and the second 1 × 1 convolution with the channel dimension is recovered.

$$s(\mathbf{F}) = \sigma(\mathbf{W_2}\delta(\mathbf{W_1 v})), \tag{4}$$

where $\mathbf{W_1} \in \mathbb{R}^{C/r \times C}$ and $\mathbf{W_2} \in \mathbb{R}^{C \times C/r}$, and σ denotes the sigmoid function.

Residual channel attention block. It is found that the residual channel attention block can provide better result than the channel attention especially when training with large datasets, such as Avenue or ShanghaiTech dataset. In the residual channel attention block, the channel attention is located after two 3 × 3 convolution layers just before the residual connection. A ReLU activation is placed between two convolutional layers as shown in Fig. 3. Given the input feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, the residual channel attention block is computed as:

$$\mathbf{F}' = \mathbf{F} \oplus (\mathbf{X} \otimes s(\mathbf{X})), \tag{5}$$

where $\mathbf{F}$ and $\mathbf{F}'$ are the input and output feature map, respectively, and $s(\mathbf{X})$ refers to the channel attention. X is obtained by:

$$\mathbf{X} = \mathbf{W}_2\delta(\mathbf{W}_1\mathbf{F}), \tag{6}$$

where δ denotes the ReLU activation function. $\mathbf{W_1}$ and $\mathbf{W_2}$ are the weight sets of the two convolutional layers.

3.5 Objective function

The goal of our network is to predict the future frame $\hat{I}_{t+1}$ from a sequence of input frames $\{I_1, I_2, ..., I_t\}$. Since each frame consists of many pixels and each pixel has an intensity, the constraints for intensity and its gradient can be the important factors in minimizing the prediction error. Thus, the similarity of all pixels in RGB space can be ensured by an intensity constraint that compares every pixel value between the predicted frame and the ground-truth frame as follows:

$$L_{int}(I, \hat{I}) = \left\| I - \hat{I} \right\|_2^2 \tag{7}$$

To deal with potential blur occurring while adopting l_2 $distance$, a gradient constraint is added to obtain a sharper video frame. The loss function computes the difference between absolute gradients along two spatial dimensions as follows:

$$\begin{aligned} L_{gra}(I, \hat{I}) = &\sum_{i,j} \left\| |\hat{I}_{i,j} - \hat{I}_{i-1,j}| - |I_{i,j} - I_{i-1,j}| \right\|_1 \\ &+ \left\| |\hat{I}_{i,j} - \hat{I}_{i,j-1}| - |I_{i,j} - I_{i,j-1}| \right\|_1 \end{aligned} \tag{8}$$

To measure Structural Similarity (SSIM), Multi-Scale Structural Similarity (MS-SSIM) is used [22, 23]. Note

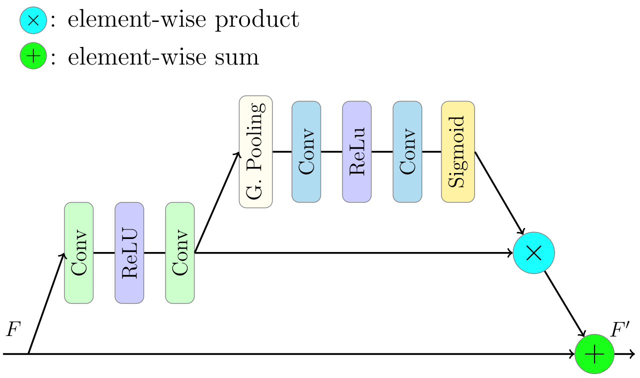


Fig. 3 Residual channel attention block. The input features are passed through two 3 × 3 convolution layers with a ReLU activation in between before the channel attention is applied

that MS-SSIM has been proposed initially for the image quality assessment at different resolutions. The combination of loss functions including intensity, gradient, and multi-scale structural similarity constraint is given as follows:

$$L_{con}(I, \hat{I}) = \alpha L_{int}(I, \hat{I}) + \beta L_{gra}(I, \hat{I}) + \gamma L_{mss}(I, \hat{I}), \quad (9)$$

where α, β, and γ are three coefficients that balance the weights between the losses.

3.6 Anomaly detection

To detect anomaly, we use anomaly score $S(t)$, which is used to measure the difference between the ground truth frame I and the predicted frame $\hat{I}$. Since the Peak Signal to Noise Ratio (PSNR) is widely used in assessing the image quality, the quality of a predicted frame is calculated as follows:

$$PSNR(I, \hat{I}) = 10log_{10}\frac{[max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=1}^{N}(I_i - \hat{I}_i)^2} \quad (10)$$

where N denotes the number of rows and columns in a frame (the number of pixels), $[max_{\hat{I}}]$ is the maximum value of $\hat{I}$. The higher value of PSNR indicates that the frame has a higher quality. In other words, the difference between the ground truth frame and the predicted frame is small.

Following [20], the PSNR of all frames in each test video is normalized to the range [0,1], and we compute the anomaly score $S(t)$ for each frame by using the following formula:

$$S(t) = \frac{PSNR_t - min(PSNR)}{max(PSNR) - min(PSNR)} \quad (11)$$

where $min(PSNR)$ and $max(PSNR)$ denote the minimum and the maximum PSNR values in the given video sequence, respectively. The anomaly score of a predicted frame indicates whether the frame is normal or abnormal with a given threshold.

4 Experimental evaluation

4.1 Datasets

Performance evaluation was carried out using three benchmark datasets such as UCSD Pedestrian dataset [26], CUHK Avenue dataset [21] and ShanghaiTech dataset [24]. Figure 4 shows the sample cases of them. In each dataset, the training set contains only normal videos, whereas the test set contains both normal and abnormal frames. In each test video, the ground truth annotation includes a binary flag per frame, indicating whether a frame contains anomaly event or not. So that, label 0 is the normal frame and label 1 is the abnormal frame.

UCSD Dataset. The UCSD dataset had two subsets, namely Ped1 and Ped2, which were recorded at two different outdoor locations. The former had a resolution of 158 × 238 and the latter a resolution of 240 × 360. Pedestrians walked across the camera. Such normal events were used for training. The abnormal events in this dataset were defined by the appearances of a car, a biker, a skater or a wheelchair. Following the work of [5, 10], Ped1 had been excluded from our experiments because of its lower resolution. Ped2 contained 16 training videos and 12 test videos, corresponding to 2550 frames for training and 2010 for testing, respectively.

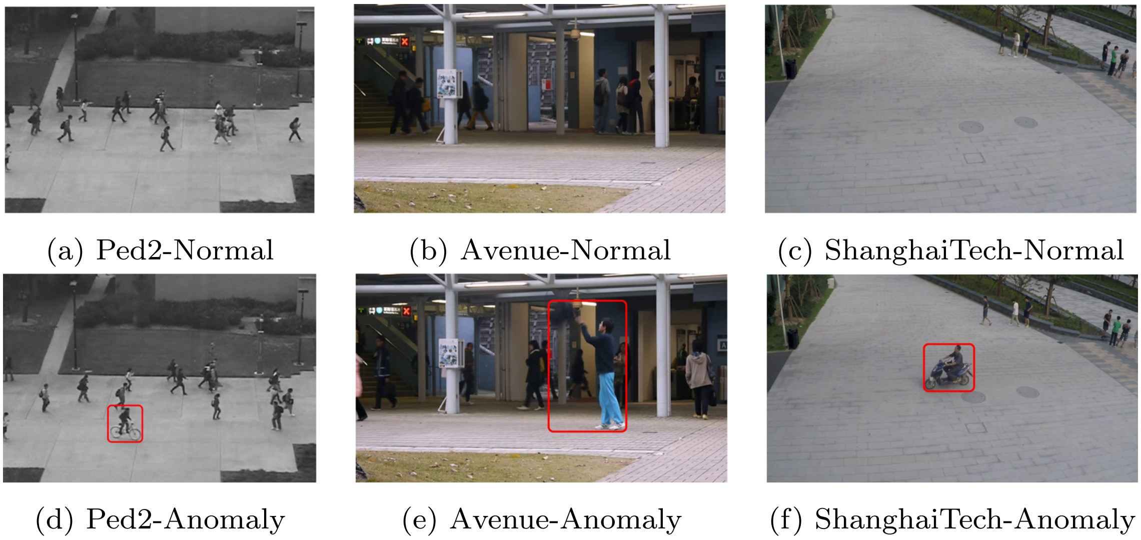
CUHK Avenue dataset. This dataset consisted of 16 training and 21 test videos, corresponding to 15,328 frames and 15,324 frames, respectively. The resolution of each video frame was 360 × 640 pixels. There were 47 abnormal events, such as throwing objects, loitering, and running across the gate.

ShanghaiTech Campus dataset. The ShanghaiTech Campus dataset was one of the most challenging datasets for video anomaly detection, containing 130 abnormal events. The dataset had 330 training and 107 test videos from 13 different scenes with various lighting conditions and camera angles. It had 317,398 frames and each frame had a resolution of 480 × 856 pixels. The dataset was split into 274,515 training frames and 42,883 test frames.

4.2 Parameter and implementation

Each video frame was resized as 224 × 288 for Ped2, 192 × 320 for CUHK Avenue, and 192 × 288 pixels for ShanghaiTech, respectively. The intensity of each frame was normalized in the range of $[-1, 1]$ before being fed into the model. The learning rate was set as 2e-4 initially and decreased to 1e-4 at epoch 60 for Ped2, 50 for Avenue, and 30 for ShanghaiTech, respectively. The Adam optimizer was adopted for training our network. To reduce computation complexity, we utilized the penultimate feature map instead of the last one extracted by a deep convolutional neural network in the encoder when training Avenue and ShanghaiTech datasets. After choosing a sequence of five video frames randomly from the training set, the first four frames among them were used as input, and the fifth frame was used as the ground truth frame. Then, the ground truth frame was compared with the predicted frame obtained from the model in calculating the anomaly score.

Evaluation metric. Following the prior work [20, 39], the frame-level area under the curve (AUC) was used in evaluating the performance of our proposed network.


(a) Ped2-Normal (b) Avenue-Normal (c) ShanghaiTech-Normal

(d) Ped2-Anomaly (e) Avenue-Anomaly (f) ShanghaiTech-Anomaly

Fig. 4 Examples of normal (the top row) and abnormal (the bottom row) frames in the UCSD Ped2, CUHK and Shanghaitech datasets, respectively. The abnormal object is denoted by a red boxes, such as a man riding a bicycle (d), throwing a bag, and riding a motorbike

The AUC was obtained by computing the area under the receiver operating characteristic (ROC) with varying threshold values for the abnormal scores. A higher AUC value indicated better anomaly detection performance.

4.3 Ablation study

4.3.1 Performance evaluation of processing units in the network

Given that our network contains three major processing units such as spatial processing, temporal processing and attention, an ablation study was carried out in order to evaluate their effectiveness in terms of performance. Table 1 shows result for the combinations of three components. First, when the temporal processing and the attention module are excluded, the network has the spatial processing unit. Secondly, the effectiveness of both spatial and temporal features without the channel attention module in the decoder is shown in the second row of Table 1. Thirdly, only spatial features are used as input of the decoder to estimate the capability of channel attention, which aims to exploit the attention across channels. Finally, the performance of the whole network is shown in the bottom row.

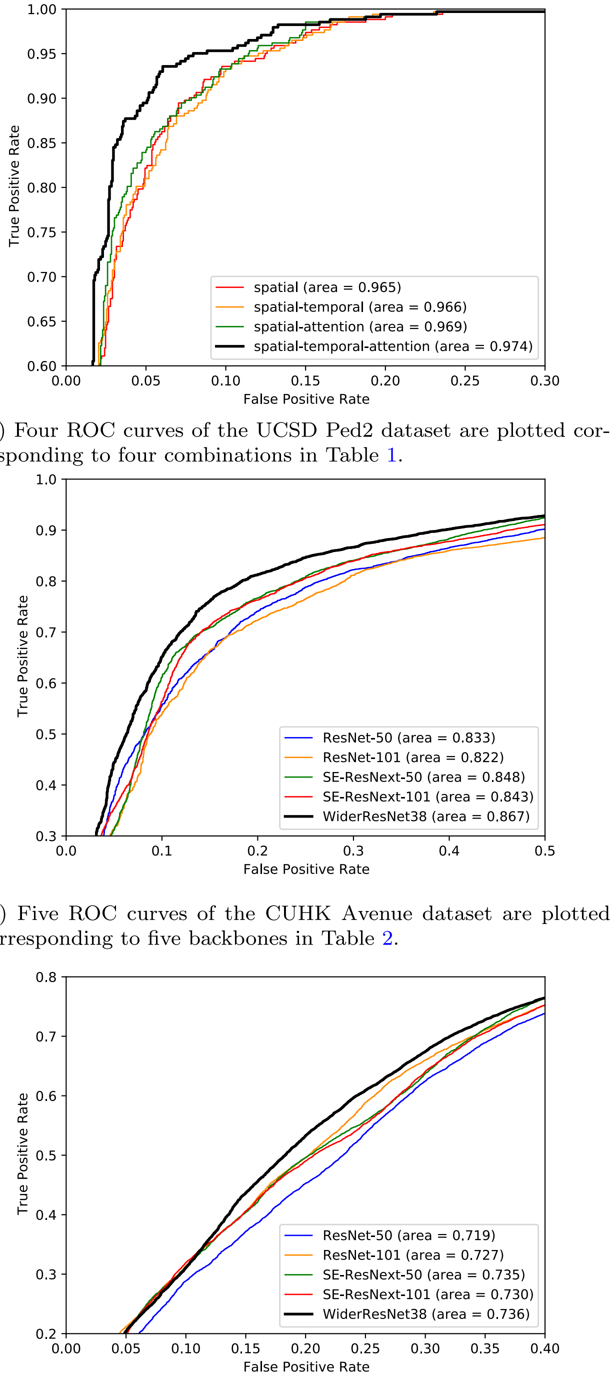
The AUC performance (%) of proposed network using WiderResnet38 [34] as backbone with different combination of componets on UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets is shown in Table 1. The performance of the baseline, which contains only spatial features, was improved by combining it with other components such as temporal and attention. For instance, the channel attention component improved the performance of the network significantly. Notice that the network using both spatio-temporal features and channel attention achieved the highest performance, reaching 97.4% for UCSD Ped2, 86.7% for CUHK Avenue and 73.6% for ShanghaiTech datasets, respectively, confirming that the combination of spatial and temporal branches provided more information for encoding the input frames, and the channel attention module played a vital role in restoring the future frame well.

In particular, the ROC curves for UCSD Ped2 dataset are shown in Fig. 5, wherein the red and orange curves denote the ROC curves of the method using spatial and spatio-temporal features, respectively. The green one denotes that

Table 1 Comparison between different processing units in the proposed network in term of AUC (%) on UCSD Ped2, CUHK Avenue and ShanghaiTech datasets. When the spatial and temporal processing unit are combined with the channel attention unit, the system performs best

Backbone	Spatial	Temporal	Attention	UCSD Ped2	CUHK Avenue	ShanghaiTech
WiderResNet	✓	–	–	96.5	84.6	71.6
	✓	✓	–	96.6	85.3	72.5
	✓	–	✓	96.9	85.7	72.4
	✓	✓	✓	**97.4**	**86.7**	**73.6**

The bold entries show the best results



(a) Four ROC curves of the UCSD Ped2 dataset are plotted corresponding to four combinations in Table 1.

(b) Five ROC curves of the CUHK Avenue dataset are plotted corresponding to five backbones in Table 2.

(c) Five ROC curves of the ShanghaiTech Campus dataset are plotted corresponding to five backbones in Table 2.

Fig. 5 Frame-level ROC curves for three benchmark datasets

of proposed method using spatial features in the encoder and channel attention modules in the decoder. The black one denotes the ROC curve of the proposed approach, which includes spatial and temporal branches in the encoder and efficient channel attention in the decoder, reaching 97.4% for UCSD Ped2 dataset.

4.3.2 Evaluation of deep convolutional neural networks as backbone

To show the effectiveness of our network architecture, this section describes the performance of the proposed network with different deep convolutional neural networks as a backbone in Table 2. The network architecture is kept unchanged and only the backbone is replaced by different deep convolutional neural networks. The proposed network with different backbones except ResNet-50 outperforms the baseline method [20] for UCSD Ped2 and ShanghaiTech datasets, suggesting that the proposed method can achieve high performance using different features that are extracted by different deep convolutional neural networks as a backbone.

For instance, our network using WiderResNet38 [34] as a backbone gives the best performance for UCSD Ped2, CUHK Avenue and ShanghaiTech datasets, achieving the AUC of 97.4%, 86.7% and 73.6%, respectively. It also achieves 96.7% using SE-ResNext-101 as a backbone [12] for UCSD Ped2, and 73.5% using SE-ResNext-50 [12] as a backbone for ShanghaiTech, respectively.

Figure 5b and c show the frame-level ROC curves using different deep neural networks as a backbone for CUHK Avenue and ShanghaiTech dataset, respectively. The blue line denotes ROC for ResNet-50 whereas the orange lines for ResNet-101. The ROC curves of the for SE-ResNext-50 and SE-ResNext-101 have green and red lines, respectively. The black line is the ROC curve for WiderResNet38 as a backbone.

4.4 Comparison with state-of-the-art

Table 3 compares our approach with the recent state-of-the-art methods for three standard anomaly datasets. These methods are categorized into three groups, such as the reconstruction-based methods, the prediction-based methods, and the hybrid methods. Among them, our method achieves the best performance for UCSD Ped2, CUHK Avenue and ShanghaiTech dataset, reaching the AUC of 97.4%, 86.7% and 73.6%, respectively. Note that the frame-level AUCs of our method are higher than that of the frame prediction-based anomaly detection baseline [20] about 2% for UCSD Ped2 and approximately 1% for CUHK Avenue and ShanghaiTech Campus datasets, suggesting that our network outperforms most of the recent anomaly detection methods in term of AUC performance. The ShanghaiTech dataset is challenging because it is a large-scale dataset, including over 270K training frames and 42K test frames. Since it contains a large amount of data having diverse types of normal and abnormal events, its performance is relatively lower than those of other datasets.

Table 2 Comparison of the proposed network with different deep convolutional neural networks as backbone in term of AUC (%). The proposed network using WiderResnet38 [34] as backbone achieves the best performance

Method	Backbone	UCSD Ped2	CUHK Avenue	ShanghaiTech
Proposed method	ResNet-50 [11]	95.1	83.3	71.9
	ResNet-101 [11]	95.8	82.2	72.7
	SE-ResNext-50 [12]	96.1	84.8	73.5
	SE-ResNext-101 [12]	96.7	84.3	73.0
	WiderResNet38 [34]	**97.4**	**86.7**	**73.6**

The bold entries show the best results

Table 3 Comparison with recent state-of-the-art methods for video anomaly detection in terms of AUC (%) on three benchmark datasets. The proposed network uses WiderResnet38 [34] as a backbone

Methods		UCSD Ped2	CUHK Avenue	ShanghaiTech
Hybrid	Tang et al. [29]	96.3	85.1	73.0
	Morais et al. [27]	-	86.3	73.4
Reconstruction	Wei et al. [30]	89.5	79.7	67.2
	Nawaratne et al. [28]	91.1	76.8	–
	Li et al. [14]	91.6	84.2	–
	Luo et al. [25]	92.2	83.5	69.6
	Gong et al. [10]	94.1	83.3	71.2
	Zhou et al. [38]	94.9	86.1	–
	Abati et al. [1]	95.4	–	72.5
	Hu et al. [13]	95.9	84.2	–
	Wu et al. [32]	92.8	85.5	–
	Li et al. [15]	92.9	83.5	–
	Fan et al. [7]	92.2	83.4	–
	Fang et al. [8]	95.6	86.3	73.2
	Chang et al. [2]	96.5	86.0	73.3
	Deepak et al. [4]	83.0	82.0	–
Prediction	Liu et al. [20]	95.4	85.1	72.8
	Li et al. [16]	96.5	84.5	–
	Lu et al. [22]	96.0	85.7	–
	Zhou et al. [39]	96.0	86.0	–
	Yang et al. [36]	95.9	85.9	73.5
	Doshi et al. [5]	97.2	86.4	70.9
	Ours	**97.4**	**86.7**	**73.6**

The bold entries show the best results

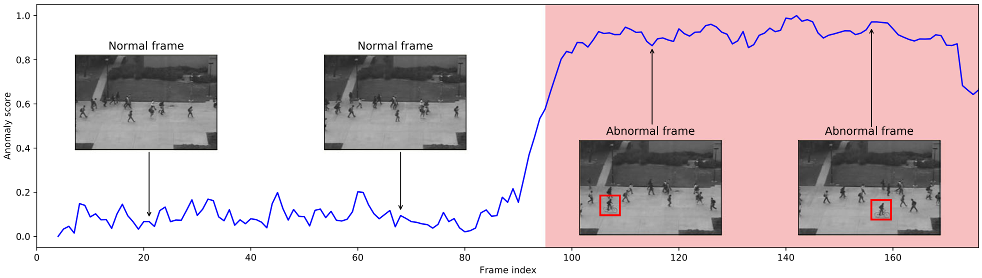


Fig. 6 Anomaly score of the test video 02 in the UCSD Ped2 dataset. The red rectangles denote the abnormal objects (riding bicycle) in the frames. Note that the anomaly score drawn as a blue line is increased as the abnormal object appears

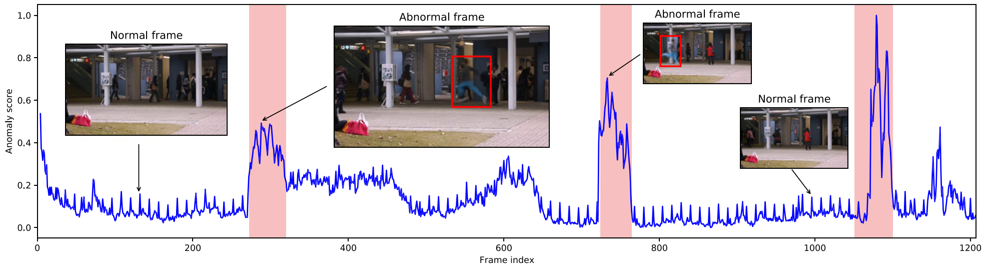


Fig. 7 Anomaly score of the test video 02 in the CUHK Avenue dataset. The three pink areas indicate the ground truths for appearing anomalous object. The red rectangles denote the abnormal running objects. The anomaly score is drawn as a blue line that is increased whenever the abnormal object is moving in front of the building. Note that the third abnormal event comes from the shaking camera

The overall shape of our proposed network is an autoencoder, consisting of encoder and decoder. In the encoder, the temporal branch is to model the temporal information by applying the effective temporal shift method which does not add any extra parameters and the shift operation is performed at zero computation [19]. As mention in Section 3.3, a 1 × 1 convolution is the main operation in the spatial branch to reduce the dimension of the aggregated features. On the other hand, channel attention uses small extra parameters and computation [31]. As shown in Fig. 3, a channel attention module includes a global average pooling, two 2D convolutions, a rectified linear unit activate and a sigmoid function. The proposed architecture appears to be an effective approach for anomaly detection in videos.

4.5 Visualization

4.5.1 Anomaly score

Figures 6, 7 and 8 show how anomaly score can be visualized along video frames for three anomaly datasets. Note that the anomaly score drawn as a blue line in each figure changes rapidly between the normal and the abnormal event, indicating that our network is able to distinguish the sporadically occurring abnormal events among the vast normal ones within a given video. Figure 6 shows how the anomaly score varies for the normal and abnormal events occurring in the test video 02 of the UCSD Ped2 dataset. The first two frames show only the walking pedestrians, whereas the remaining two frames contain a bicycle rider among these pedestrians. Notice that the anomaly score increases dramatically when the rider appears within the frame and the score maintains high level until he disappears.

Figure 7 visualizes how the anomaly score changes as a running man appears in front of a building from the test video 02 of the CUHK Avenue dataset. Three abnormal events are shown: The first two abnormal events record the man running, and the third event come from the shacking of the camera. The anomaly score rises steeply when the man appears and then decrease sharply when he steps out of the frame. A noticeable fact is that the anomaly scores of the third event shows the highest score presumably because the camera shake event affects the whole frame.

The anomaly score and some key frames of the test video 01_0063 of the ShanghaiTech dataset are visualized in Fig. 8. Two normal frames contain a few pedestrians on the walkway, while the abnormal event has a bicycle rider. The anomaly score increases rapidly when a bicyclist comes into the scene, whereas the score decreases as the rider disappears, confirming that our network is able to distinguish the abnormal frame from the vast normal frames.

4.5.2 Network visualization

In this section, the visualizations[2] of UCSD Ped2 and ShanghaiTech Campus datasets are shown in Figs. 9 and 10, respectively. Each figure visualizes a sample of normal event on the left column and an abnormal event on the right column. The ground truth frames of events are shown on the first row of each figure. The extracted features obtained from DCNN are fed to the spatial and temporal branches. The spatio-temporal features are visualized on the second row. In addition, channel attention [31] is applied to focus on some objects among others, visualizing as the attention maps on the third row. Given that prediction error can be measured as the difference between the predicted frame and its ground truth frame, our network is designed to produce a smaller prediction error for the normal frame, whereas a larger prediction error for the abnormal frame. Following [36, 39], the prediction errors for UCSD Ped2, and ShanghaiTech datasets are visualized at the last row of Figs. 9 and 10, respectively.

In Fig. 9, the ground truth samples of a normal and abnormal events (a, b) are shown on the first row and the corresponding spatio-temporal maps (c, d) are visualized on the second row. In the attention map (f), the cyclist appears to be salient among the pedestrians. On the bottom row, the prediction error around the cyclist is bigger than those

[2] The demo video can be found at https://youtu.be/XOzXwKVKX-Y

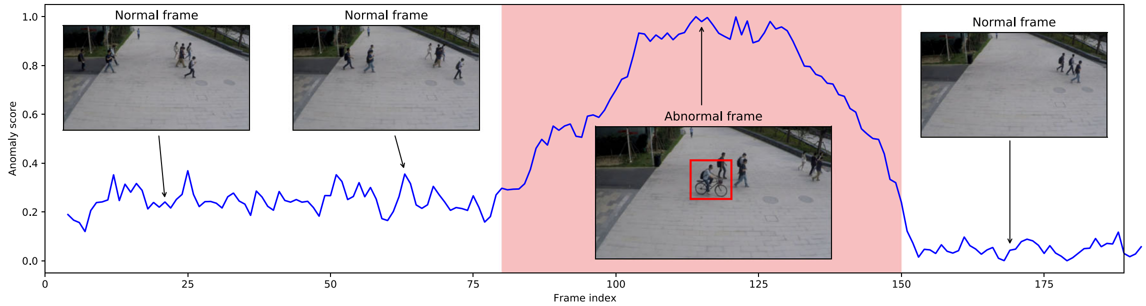
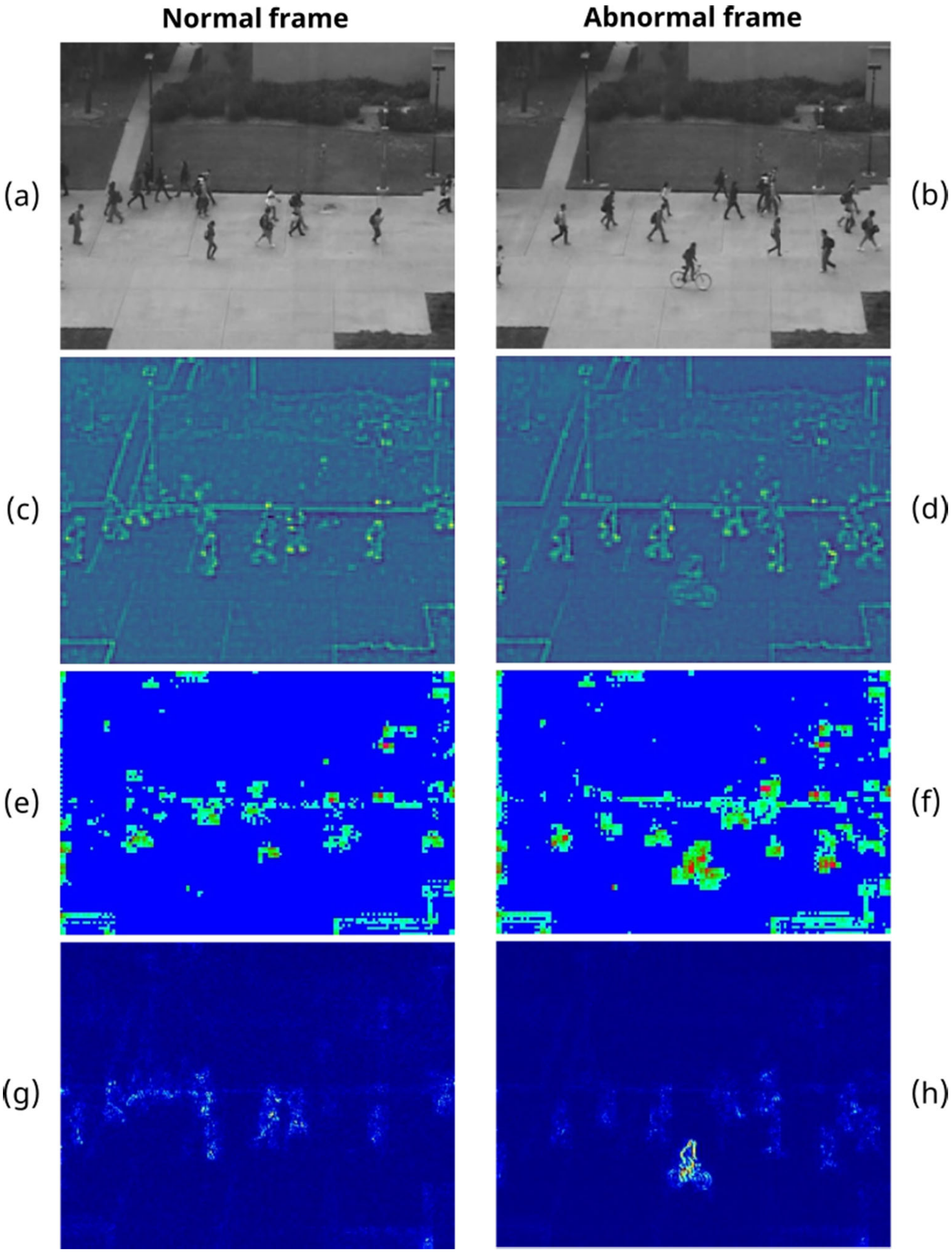


Fig. 8 Anomaly score of the test video 01_0063 in the ShanghaiTech dataset. The pink area indicates the ground truth. The red rectangle denotes the abnormal riding object

Fig. 9 Network visualization for the UCSD Ped2 dataset. The normal and abnormal samples are shown on the left and the right column. From top to bottom, the ground truth frames (a, b), the spatio-temporal maps (c, d), the attention maps (e, f), and the prediction errors (g, h) are shown, respectively. Note that the attention map has half the resolution of input video



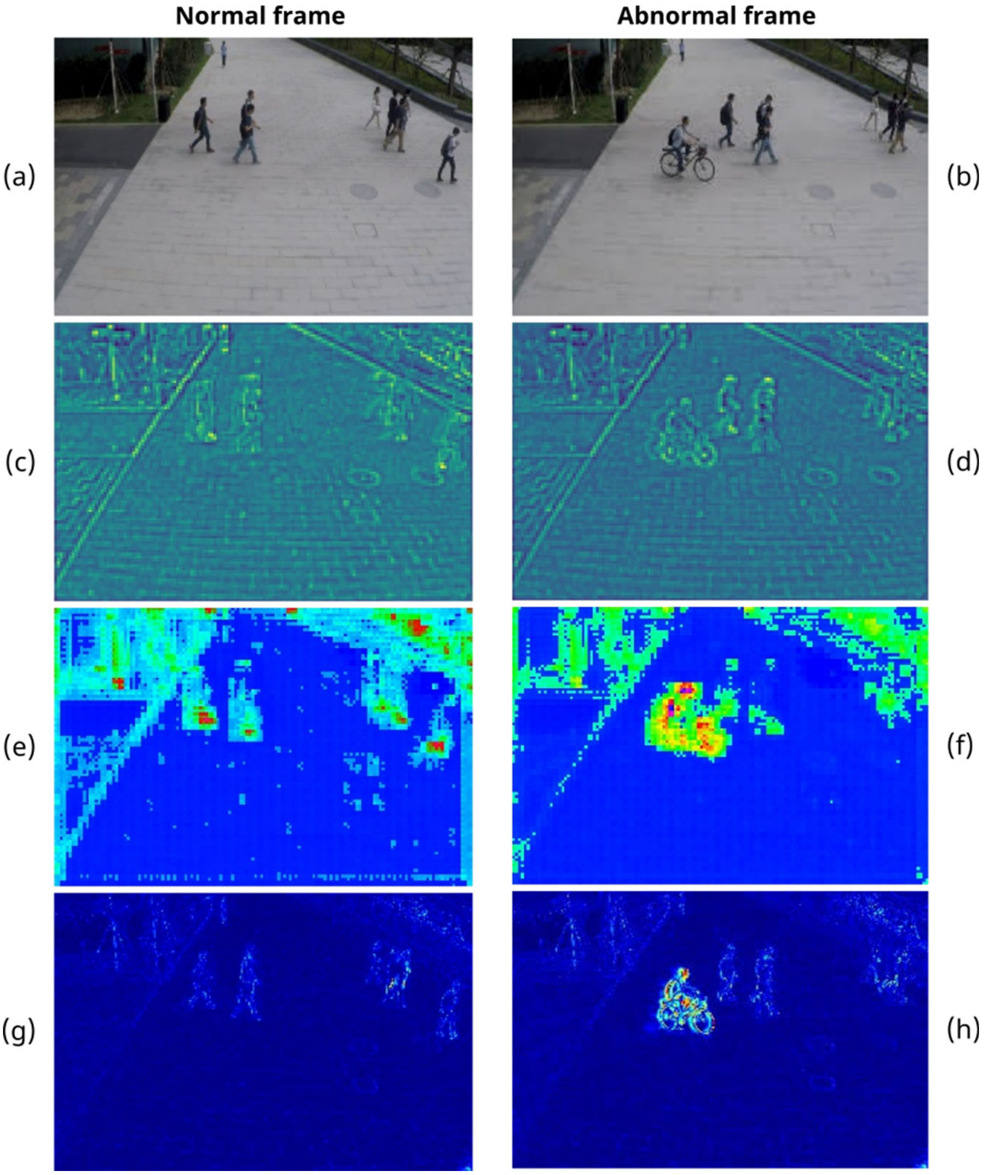


Fig. 10 Network visualization for the ShanghaiTech Campus dataset. The normal case is shown on the left column and the abnormal one is shown on the right column. From top to bottom, the ground truth frames (a, b), the spatio-temporal maps (c, d), the attention maps (e, f) and the prediction errors (g, h) are shown, respectively. Note that the attention map has half the resolution of input video

around pedestrians in the abnormal case (h) since the model has been trained with normal frames.

A similar observation can be made for ShanghaiTech dataset as shown in Fig. 10, wherein the cyclist is also seen as an abnormal object. The spatio-temporal maps (c, d) corresponding to the ground truth frames (a, b) of the normal and abnormal events are shown in the second row. On the third row, attention of the network is distributed within the normal frame, whereas it is focused around the cyclist within the abnormal frame. The difference between the ground truth and the predicted frames is illustrated as prediction error in the bottom row. Similar to the above case, the prediction error of the normal frame (g) is minimal, whereas the riding cyclist as an abnormal event produces a large prediction error around him as shown in (h).

5 Conclusion and future work

This study presents a new video anomaly detection framework that has an attention-based residual autoencoder architecture. The proposed network is based on unsupervised learning and it exploits both spatial and temporal information in a unified network. The temporal shift is developed for the effective temporal feature extraction. In addition, the channel attention mechanism is utilized to exploit the channel relationship of features, which significantly helps the model learn more effectively. Experiments on three anomaly benchmark datasets show that our network outperforms the state-of-the-art methods. The ablation study shows that not only the spatio-temporal circuits in encoder but also the cascade type application of channel attention in decoder have been very effective in improving the system performance. Moreover, the proposed network architecture works well in 2D data and may be generalizable to 3D data for real-world engineering applications [17, 18]. We look forward to applying this framework for the practical surveillance systems.

Acknowledgements We would like to thank to reviewers for their valuable suggestions to improve our manuscript. This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP), grant funded by the Korea government (MSIT) (No.2019-0-00231) as well as by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540).

References

1. Abati D, Porrello A, Calderara S, Cucchiara R (2019) Latent space autoregression for novelty detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 481–490
2. Chang Y, Tu Z, Xie W, Yuan J (2020) Clustering driven deep autoencoder for video anomaly detection. In: European Conference on Computer Vision, Springer, pp 329–345
3. Chang Y, Tu Z, Xie W, Luo B, Zhang S, Sui H, Yuan J (2022) Video anomaly detection with spatio-temporal dissociation. Pattern Recogn 122:108213
4. Deepak K, Chandrakala S, Mohan CK (2021) Residual spatiotemporal autoencoder for unsupervised video anomaly detection. SIViP 15(1):215–222
5. Doshi K, Yilmaz Y (2021) Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. Pattern Recogn 114:107865
6. Doshi K, Yilmaz Y (2022) Rethinking video anomaly detection-a continual learning approach. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3961–3970
7. Fan Y, Wen G, Li D, Qiu S, Levine MD, Xiao F (2020) Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Comput Vis Image Underst 195:102920
8. Fang Z, Zhou JT, Xiao Y, Li Y, Yang F (2020) Multi-encoder towards effective anomaly detection in videos. IEEE Transactions on Multimedia
9. Georgescu MI, Barbalau A, Ionescu RT, Khan FS, Popescu M, Shah M (2021) Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12742–12752
10. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Hengel Avd (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1705–1714
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
12. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
13. Hu J, Zhu E, Wang S, Liu X, Guo X, Yin J (2019) An efficient and robust unsupervised anomaly detection method using ensemble random projection in surveillance videos. Sensors 19(19):4145
14. Li N, Chang F (2019) Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder. Neurocomputing 369:92–105
15. Li N, Chang F, Liu C (2020) Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Transactions on Multimedia 23:203–215
16. Li Y, Cai Y, Liu J, Lang S, Zhang X (2019) Spatio-temporal unity networking for video anomaly detection. IEEE Access 7:172425–172432
17. Liang Y, He F, Zeng X (2020) 3d mesh simplification with feature preservation based on whale optimization algorithm and differential evolution. Integrated Computer-Aided Engineering 27(4):417–435
18. Liang Y, He F, Zeng X, Luo J (2022) An improved loop subdivision to coordinate the smoothness and the number of faces via multi-objective optimization. Integrated Computer-Aided Engineering, pp 23–41
19. Lin J, Gan C, Han S (2019) Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7083–7093
20. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6536–6545
21. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision, pp 2720–2727
22. Lu Y, Kumar KM, shahabeddin Nabavi S, Wang Y (2019) Future frame prediction using convolutional vrnn for anomaly detection. In: 2019 16Th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, pp 1–8
23. Lu Y, Yu F, Reddy MKK, Wang Y (2020) Few-shot scene-adaptive anomaly detection. In: European conference on computer vision, Springer, pp 125–141
24. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE international conference on computer vision, pp 341–349
25. Luo W, Liu W, Lian D, Tang J, Duan L, Peng X, Gao S (2019) Video anomaly detection with sparse coding inspired deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence
26. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 1975–1981
27. Morais R, Le V, Tran T, Saha B, Mansour M, Venkatesh S (2019) Learning regularity in skeleton trajectories for anomaly detection in videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11996–12004
28. Nawaratne R, Alahakoon D, De Silva D, Yu X (2019) Spatiotemporal anomaly detection using deep learning for real-time

video surveillance. IEEE Transactions on Industrial Informatics 16(1):393–402

29. Tang Y, Zhao L, Zhang S, Gong C, Li G, Yang J (2020) Integrating prediction and reconstruction for anomaly detection. Pattern Recogn Lett 129:123–130
30. Wei H, Li K, Li H, Lyu Y, Hu X (2019) Detecting video anomaly with a stacked convolutional lstm framework. In: International conference on computer vision systems, Springer, pp 330–342
31. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
32. Wu P, Liu J, Li M, Sun Y, Shen F (2020) Fast sparse coding networks for anomaly detection in videos. Pattern Recogn 107:107515
33. Wu Y, He F, Zhang D, Li X (2015) Service-oriented feature-based data exchange for cloud-based design and manufacturing. IEEE Transactions on Services Computing 11(2):341–353
34. Wu Z, Shen C, Van Den Hengel A (2019) Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recogn 90:119–133
35. Xingjian S, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp 802–810
36. Yang Y, Zhan D, Yang F, Zhou XD, Yan Y, Wang Y (2020) Improving video anomaly detection performance with patch-level loss and segmentation map. In: 2020 IEEE 6th international conference on computer and communications (ICCC), IEEE, pp 1832–1839
37. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
38. Zhou JT, Du J, Zhu H, Peng X, Liu Y, Goh RSM (2019) Anomalynet: an anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security 14(10):2537–2550
39. Zhou JT, Zhang L, Fang Z, Du J, Peng X, Xiao Y (2019) Attention-driven loss for anomaly detection in video surveillance. IEEE Trans Circuits Syst Video Technol 30(12):4639–4647

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Viet-Tuan Le received the M.Sc. degree in computer science from the University of Science, Vietnam National University, Ho Chi Minh City, Vietnam, in 2012. He is now a Ph.D. candidate in Computer Science at the Computer Engineering Department, Sejong University, Korea. His research interests include action recognition and anomaly detection.

Yong-Guk Kim received the B.S. and M.S. degrees in Electrical Engineering from Korea University, Seoul, Korea, in 1982 and 1984, respectively. After several years of works at LG Electronics and Korea Telecom (KT) as a researcher, he studied in University of Cambridge, UK and received the Ph.D. focusing on computational vision in 1997. From 1995 to 1996, he was a Research Fellow with Helmholtz Robotics Institute, Utrecht, the Netherland, and from 1998 to 2001, was a Research Associate with Smith-Kettlewell Vision Institute, San Francisco, USA.

He has been in the Computer Engineering Department, Sejong University, Seoul, Korea since 2001 and is a full Professor. His posts in the university include director of start-up incubator and Dean of international affairs. His research areas are facial expression recognition, driver drowsiness detection, anomaly detection, hand pose estimation and vision-based autonomous navigation of the drone.