An Attention-based U-Net Network for Anomaly Detection in Crowded Scenes

Jinpeng Fang, Xinfeng Zhang*, Baoqing Yang, Shuhan Chen, Bin Li
College of Information Engineering, Yangzhou University
Yangzhou, China
zhangxf@yzu.edu.cn

2022 14th International Conference on Computer Research and Development (ICCRD) | 978-1-7281-7721-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCRD54409.2022.9730481

***Abstract*—Anomaly detection in surveillance video is of great significance for public safety. Deep autoencoder has been widely used in anomaly detection. Because of its good generalization ability, sometimes abnormal samples can still be reconstructed very well. Some scholars use memory module constructed by using the normal samples to reconstruct the test samples. The memory items need to be retrieved and updated during the training and testing process, hence more memory space is required to store memory module, which greatly increases the training and test costs. We tackle the problem of the excessive generalization ability of autoencoder from a new perspective. We introduce an attention mechanism to propose an attention-based U-Net network to detect anomalies. The network adds an attention module before the skip connection of U-Net network, so that the model pays more attention to the foreground targets. During the training process, the normal foreground targets are learned more fully. Therefore, in the test, the proposed method can achieve more accurate prediction of normal targets that appear frequently, so that the rare abnormal targets can be highlighted because of the large prediction errors. We conduct experiments on real surveillance videos of UCSD Ped1 and ShanghaiTech datasets, and the experimental results demonstrate that the proposed method is an efficient model.**

Keywords—anomaly detection, video surveillance, deep autoencoder, attention mechanism.

I. INTRODUCTION

Nowadays, a large number of surveillance devices are installed in various public places, and the generated massive surveillance videos need to be processed in real-time to meet people's safety requirements. In surveillance videos, people usually do not care about the recurring normal events, and the rare abnormal events should be focused on, such as a motor vehicle appears on the sidewalk. Detecting abnormal behaviors in crowded public places is very challenging for the main following reasons: Public places where people gather are disturbed by environmental clutter, interlacing, dynamic occlusion, etc., which makes it infeasible to realize anomaly detection by analyzing each individual in such crowded scene. Secondly, abnormal behaviors rarely occur, and their patterns are variable. Obviously, it is impractical to construct a complete abnormal model with a small number of abnormal samples [6]. Since it is relatively easy to obtain normal surveillance videos, abnormal behavior detection is usually regarded as an unsupervised problem that only uses normal samples for training [1][3].

*Corresponding author E-mail address: zhangxf@yzu.edu.cn (Xinfeng Zhang).

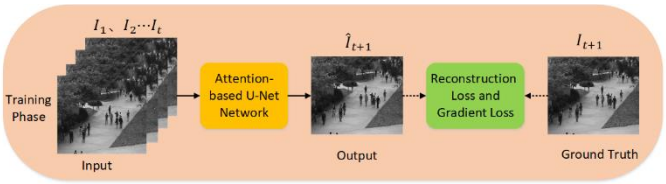
Because the anomaly detection problem has important practical and theoretical significance, it has attracted researchers to invest a lot of energy in related research. Generally, unsupervised anomaly detection methods use the prior information in the normal videos to train the model. Events that cannot be described by the normal model during the test are considered anomalies [4], which are usually judged based on reconstruction or prediction errors. Compared with handcrafted features, deep learning can provide more descriptive features. Hasan et al. [8] use a convolutional autoencoder (Con-AE) to model the normal continuous multiple video frames and then use the normal model to reconstruct these test frames, and those with large reconstruction errors are regarded as anomalies. Deep convolutional autoencoder has strong generalization capabilities [2]. The autoencoder trained on normal samples can sometimes reconstruct abnormal samples well. Therefore, it is difficult to obtain ideal anomaly detection results by distinguishing the size of the reconstruction error. Gong et al. [2] propose a memory-based autoencoder, which records normal patterns in the memory module during training and use the coding features of the test sample to retrieve the most relevant memory items in the memory module during testing, and then use these memory items to reconstruct test sample to alleviate the problem of the excessive generalization ability of autoencoder. Park et al. [4] propose an update method of memory module, which uses the coding features of sample to retrieve memory items and update them during training and testing. These memory module-based methods need to constantly retrieve and update memory items during training and testing, which greatly increases time and modeling cost, and require more space to store memory module.

The above methods perform the same reconstruction or prediction operation on both the foreground and background. In fact, the foreground targets need to be paid more attention. For this reason, we propose an attention-based U-Net network for anomaly detection. The network utilizes consecutive t frames as input to predict the $t+1$-th frame, and detects anomaly based on the difference between the predicted frame and the real frame. Before each skip connection, the output features of the encoder are adjusted by the attention module, and then merged with the up-sampled feature maps of the decoder. Attention mechanism makes the model more focused on the foreground targets, so that the normal foreground targets learn more fully during the training process. Therefore, the normal targets can be predicted more accurately in the test, so that the abnormal targets with less occurrences can be highlighted because of the large prediction errors. The proposed method tackles the problem of the

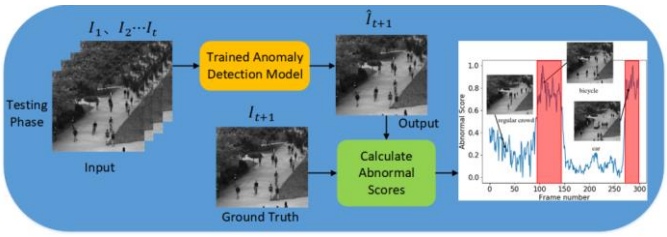
excessive generalization ability of autoencoder by introducing an attention mechanism, and avoids the construction of memory module with high computational cost. Through testing on several real scene surveillance videos, the experimental results demonstrate the proposed method is effective and efficient.

II. Our Approach

In surveillance videos, because the background is almost unchanged, it is necessary to focus on the foreground targets. To this end, we introduce an attention mechanism and propose an attention-based U-Net anomaly detection model. The training and testing process of the model is shown in Fig. 1. In this study, we use U-Net Convolutional Neural Network [10] with an attention module to construct an encoder-decoder. By introducing the attention mechanism, the attention of the model on the foreground targets can be improved, and the foreground targets can be learned more fully [9]. In the training phase, the consecutive t frames I_1、$I_2 \cdots I_t$ are fed into the encoder-decoder to predict the $t+1$-th frame I_{t+1}. In the test phase, the continuous t frames in the test videos are fed into the trained encoder-decoder, and the anomaly is judged according to the difference between the predicted frame $\hat{I}_{t+1}$ and the real frame I_{t+1}.



(a)



(b)

Fig. 1. Overall framework of anomaly detection. (a) The attention-based anomaly detection model is trained with reconstruction loss and gradient loss. (b) The trained model is used to test and the abnormal score of the video frame is calculated to judge whether the frame is abnormal.

A. Attention-based U-Net Network Architecture

The proposed anomaly detection encoder-decoder is implemented by U-Net network with an attention module, and the network structure is shown in Fig. 2. The path on the left side of the network corresponds to the encoder; the path on the right side of the network corresponds to the decoder. The encoder extracts features by reducing the spatial resolution of the $C_{input} \times H_{input} \times W_{input}$ (C_{input}, H_{input} and W_{input} represent the number of channels, height and width respectively) matrix composed of continuous t frames I_1、$I_2 \cdots I_t$ (the number of channels per frame is 3) layer by layer. Each layer consists of two 3×3 convolutions and utilizes max-pooling for down-sampling. The decoder increases the spatial resolution of the feature map layer by layer to predict the I_{t+1} frame, and uses the nearest neighbor interpolation algorithm to conduct up-sampling after performing 3×3 convolution twice in each layer. The last layer performs two 3×3 convolutions and one 1×1 convolution operation to combine the features to obtain the predicted frame $\hat{I}_{t+1}$ (the number of channels is 3). In the skip connection, the feature map of each resolution in the encoder and the corresponding feature map in the decoder are sent to the attention module [9], and then the output of the attention module is concatenated with the corresponding feature map in the decoder.

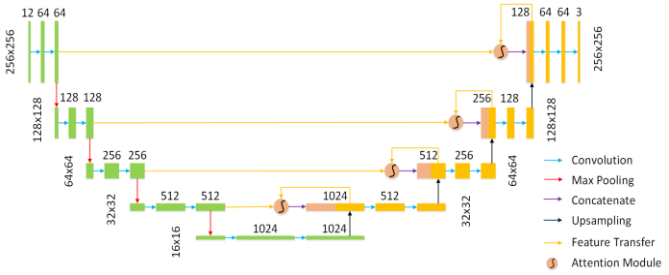


Fig. 2. Attention-based U-Net network architecture.

B. Attention Module

The low-level features extracted by using the encoder in U-Net contain a lot of redundant information. For this reason, when the feature map of each layer of the encoder is concatenated to the up-sampling layer of the decoder, we adopt an attention module [9], the structure is shown in Fig. 3.

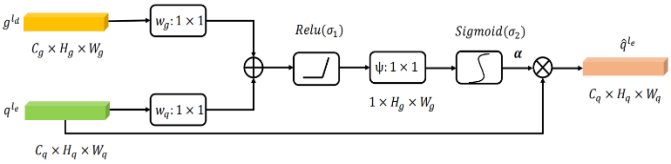


Fig. 3. Schematic diagram of attention module.

The feature map of the l_e layer in the encoder is denoted as q^{l_e}, and the size of q^{l_e} is $C_q \times H_q \times W_q$, which represent the number of channels, height and width of the feature map, respectively. The feature vector of the i-th spatial position in q^{l_e} is denoted as $q_i^{l_e}$. In the decoder, the feature map of layer $l_d - 1$ after up-sampling is denoted as g^{l_d}, and the size of g^{l_d} is $C_g \times H_g \times W_g$, which represent the number of channels, height and width of the feature map, respectively. The feature vector of the i-th spatial position in g^{l_d} is denoted as $g_i^{l_d}$. Since the feature vectors activated in the high-level feature map are usually concentrated in the salient target regions [18], the high-level feature map g^{l_d} can be used to supervise the low-level feature map q^{l_e} to make the model pay more attention to the foreground targets. Here, the weight of each spatial location in q^{l_e}, that is the attention coefficient α, is obtained according to the mapping relationship between each spatial position in g^{l_d} and the corresponding position in q^{l_e}. The attention coefficient matrix α is obtained by supervising q^{l_e} tends with g^{l_d}. The obtained attention coefficient matrix α is larger in the foreground target regions, and smaller in the unrelated background regions. Through multiplication of q^{l_e} and the attention coefficient α, the background features are diluted and target regions are highlighted.

In the attention module, the linear transformation of feature is realized by 1×1 convolution operation on q^{l_e} and g^{l_d} respectively. The two elements of corresponding positions in feature map are added, and then sent to the ReLU activation layer to realize the nonlinear transformation. The calculation formula of the ReLU activation function is as follows:

$$\sigma_1\left(q_{i,c}^{l_e}\right) = max\left(0, q_{i,c}^{l_e}\right) \tag{1}$$

where i and c represent the spatial position and channel position in the feature map, respectively. Then, a 1×1 convolution operation is performed on the obtained response value vector to compress the feature channels into a single channel. The formula is as follows:

$$f_{att} = \phi^T\left(\sigma_1\left(W_q^T q_i^{l_e} + W_g^T g_i^{l_d} + b_g\right)\right) + b_\phi \tag{2}$$

where W_q, W_g and ϕ represent the slopes of the linear transformations and b_g and b_ϕ are the bias terms. The single-channel feature map is input into the Sigmoid activation function. The formula of Sigmoid activation function is as follows:

$$\sigma_2\left(q_{i,c}^{l_e}\right) = \frac{1}{1+exp\left(-q_{i,c}^{l_e}\right)} \tag{3}$$

At this time, each feature vector $q_i^{l_e}$ obtains its corresponding attention coefficient $\alpha_i \in [0,\ 1]$. The formula of attention coefficient is as follows:

$$\alpha_i = \sigma_2\left(f_{att}\left(q_i^{l_e}, g_i^{l_d}; \theta_{att}\right)\right) \tag{4}$$

where θ_{att} is the parameters of the linear transformations. The attention coefficient matrix α is multiplied with the encoder feature map q^{l_e} to obtain the output $\hat{q}^{l_e}$ of the attention module. The output $\hat{q}^{l_e}$ is calculated as follows:

$$\hat{q}^{l_e} = q^{l_e} \cdot \alpha \tag{5}$$

Compared with the low-level features directly obtained by the encoder, $\hat{q}^{l_e}$ highlights the foreground targets and suppresses the feature activations of irrelevant background regions.

C. Loss Function

We apply the loss function consist of reconstruction loss L_{rec} and gradient loss L_{gra} to train the attention-based model. The formula of loss function is as follows:

$$L = \alpha_{rec} L_{rec} + \alpha_{gra} L_{gra} \tag{6}$$

where α_{rec} and α_{gra} are hyperparameters for adjusting the balance of reconstruction loss L_{rec} and gradient loss L_{gra}. The reconstruction loss L_{rec} describes the deviation between the pixel intensity of the predicted frame $\hat{I}$ and the real frame I. The formula of calculating the deviation by using L2 norm is as follows:

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

By penalizing the pixel intensity, the predicted frame $\hat{I}$ can be closer to the real frame I [13]. The gradient loss L_{gra} describes the difference between the image gradient of the predicted frame $\hat{I}$ and the real frame I. The formula of gradient loss is as follows:

$$L_{gra}(I, \hat{I}) = \sum_{r,c} \left\| |\hat{I}_{r,c} - \hat{I}_{r-1,c}| - |I_{r,c} - I_{r-1,c}| \right\|_1 + \left\| |\hat{I}_{r,c} - \hat{I}_{r,c-1}| - |I_{r,c} - I_{r,c-1}| \right\|_1 \tag{8}$$

where r and c represent the spatial index of the pixel. By penalizing the gradient, the predicted frame $\hat{I}$ can retain more detailed information of the real frame I [13].

D. Abnormal Judgment

In the test process, continuous t frames are fed into the trained encoder-decoder to predict the $t+1$ -th frame. Anomaly is judged by calculating the difference between the predicted frame $\hat{I}_{n+t}$ and the real frame I_{n+t}, where n represents the number of tests, $n \in [1, N-t]$, and N is the number of frames of the test video. First, we use the peak signal-to-noise ratio (PSNR) of image quality evaluation index to measure the difference between the predicted frame $\hat{I}_{n+t}$ and the real frame I_{n+t} [13]. The formula of PSNR is as follows:

$$P\left(\hat{I}_{n+t}, I_{n+t}\right) = 10\, log_{10} \frac{\max(\hat{I}_{n+t})^2}{\|\hat{I}_{n+t} - I_{n+t}\|_2^2 / M} \tag{9}$$

where M is the number of pixels in each frame. We calculate the anomaly score S_{n+t} of the current frame by normalizing PSNR. The formula is as follows [4][5]:

$$S_{n+t} = 1 - \frac{P(\hat{I}_{n+t}, I_{n+t}) - \min(\mathbf{P})}{\max(\mathbf{P}) - \min(\mathbf{P})} \tag{10}$$

where $\mathbf{P} = \left[P\left(\hat{I}_{1+t}, I_{1+t}\right), P\left(\hat{I}_{2+t}, I_{2+t}\right), \cdots, P\left(\hat{I}_N, I_N\right)\right]$, $S_{n+t} \in [0,1]$. When the video frame is normal, the anomaly score S_{n+t} is usually low; when the video frame is abnormal, the anomaly score S_{n+t} is usually high. Therefore, the anomaly score S_{n+t} can be used to distinguish abnormal cases.

III. Experiment

We test the proposed attention-based U-Net anomaly detection model on two public real scene datasets: UCSD Ped1 dataset [11] and ShanghaiTech dataset [7]. The resolution of all video frames is uniformly adjusted to 256×256, the pixel value is standardized to the interval between 0 and 1, and the number t of consecutive video frames is set to 4, that is, the 5-th frame is predicted by using 4 consecutive frames.

A. Evaluation Metric

In the work of anomaly detection, the common evaluation metric is the area under the ROC curve, i.e., AUC. The curve (ROC) is formed by gradually changing the threshold above which the test sample is judged to be abnormal [11][12]. The

vertical axis of the ROC curve is "True Positive Rate" (*TPR*), as shown in formula (11):

$$TPR = \frac{True\ Positive\ Frames}{Positive\ Frames} \tag{11}$$

where *True Positive Frames* is the number of abnormal frames in the ground truth that are correctly detected as abnormal frames and *Positive Frames* is the number of frames of all abnormal frames in ground truth. The horizontal axis of the ROC curve is "False Positive Rate" (*FPR*), as shown in formula (12):

$$FPR = \frac{False\ Positive\ Frames}{Negative\ Frames} \tag{12}$$

where *False Positive Frames* is the number of normal frames in the ground truth that are falsely detected as abnormal frames and *Negative Frames* is the number of frames of all normal frames in ground truth. We use frame-level AUC [7] to evaluate the performance of the methods. The larger the AUC value, the better the anomaly detection performance of the method.

B. Anomaly Detection on UCSD Ped1 Dataset

The UCSD Ped1 dataset [11] records the scene of pedestrians moving away from and approaching the camera. In UCSD Ped1 dataset, the training set contains 34 videos, and the test set contains 36 videos. Each video has about 200 frames with a resolution of 238×158. The test set includes 40 abnormal events.

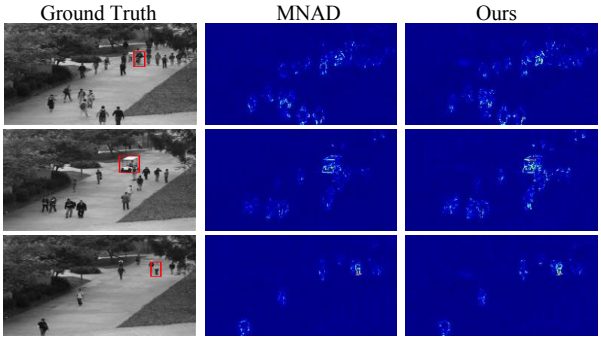


Fig. 4. Some experimental results of our method and MNAD on UCSD Ped1 dataset. The first column shows the abnormal video frames, and the second and third columns show the heatmaps obtained by visualizing prediction errors of our method and MNAD.

In order to evaluate the proposed method, we compare the AUC of the proposed method with MPPCA [14], MPPC+SFA [11], MDT [11], Unmasking [15], Conv-AE [8], ConvLSTM-AE [17], Stacked RNN [7], Frame-Pred [5], MemAE [2] and MNAD [4] on the UCSD Ped1 dataset. The results are shown in TABLE I. AUC of the proposed method is 84.9% on the UCSD Ped1 dataset. It can be seen that the proposed method achieves the best performance on UCSD Ped1. Some heatmaps of normalized prediction errors of the proposed method and MNAD for anomaly detection on the UCSD Ped1 test set are shown in Fig. 4. It can be seen that the heatmaps obtained by our method are brighter in the abnormal regions, indicating that the difference between normal cases and abnormal cases is more significant. This is because the attention-based U-Net network is more effective for learning normal foreground targets. In the test, the normal foreground targets can be predicted more accurately, so that the rare abnormal targets can be highlighted because of the large prediction errors. A large difference between prediction errors of normal cases and those of abnormal cases is conducive to more accurate anomaly detection and abnormal regions location.

TABLE I. AUC OF DIFFERENT METHODS ON THE UCSD PED1 DATASET

Method	AUC
MPPCA	59.0%
MPPC+SFA	66.8%
MDT	81.8%
Unmasking	68.4%
Conv-AE	75.0%
ConvLSTM-AE	75.5%
Frame-Pred	83.1%
MNAD	81.1%
Ours	**84.9%**

C. Anomaly Detection on ShanghaiTech Dataset

The ShanghaiTech dataset [7] contains a large number of different campus scenes, and the number is up to 13. The training set contains 330 videos and the test set contains 107 videos.

Some experimental results of the proposed method on the test set are shown in Fig. 5. It can be seen that the highlighted regions in the second row correspond to the abnormal regions marked in the first row, indicating that the prediction errors of abnormal regions are large, and abnormal regions can be accurately localized.

Fig. 5. Some test results on ShanghaiTech dataset. The first row shows the abnormal video frames, and the abnormal behaviors are a person being pushed to the ground, a running pepple, several people pushing and fighting with each other, and an electric vehicle driving on the sidewalk. The second row shows the heatmaps obtained by visualizing prediction errors. The brighter the target area, the greater the difference.

In order to evaluate the proposed method, we compare the AUC of the proposed method with Conv-AE [8], Stacked RNN [7], MemAE-nonSpar [2], Frame-Pred [5], MemAE [2] and MNAD [4] on the ShanghaiTech dataset. TABLE II lists the AUC values of various methods. The Frame-Pred method uses a generative adversarial network structure, while introducing optical flow loss, and achieves the best performance with an AUC of 72.8%. Since the Frame-Pred method uses FlowNet [16] to estimate optical flow, it inevitably brings a large amount of calculation. In addition, the training cost of the generative adversarial network is high. The proposed method is an attention-based encoder-decoder, and achieves comparable

performance to the Frame-Pred method with an AUC of 71.0%, which is better than 70.5% of MNAD. MNAD uses several memory items to record prototypical patterns of normal activities, but the ShanghaiTech dataset contains a variety of scenarios. These memory items are not enough to hold the variable patterns of normal activities, so anomaly detection performance is slightly insufficient.

TABLE II. AUC OF DIFFERENT METHODS ON THE SHANGHAITECH DATASET

Method	AUC
Conv-AE	60.9%
TSC	67.9%
StackRNN	68.0%
MemAE-nonSpar	68.8%
Frame-Pred	**72.8%**
MemAE	71.2%
MNAD	70.5%
Ours	71.0%

D. Ablation Study

In the previous section, we perform numerous quantitative comparison experiments and visualization to prove our method's effectiveness compared with the previous state-of-the-art methods. In this section, we perform some further experiments to verify the importance of attention module in our method.

To verify the effectiveness of the attention module in our method, we compare our method with a naive baseline model (based on the U-Net model without attention module) on UCSD Ped1 and ShanghaiTech datasets. From TABLE III, we can see that our method with attention module achieves a higher AUC than that without attention module.

TABLE III. THE AUC OF THE PROPOSED METHOD WITH U-NET WITHOUT ATTENTION MODULE ON UCSD PED1 AND SHANGHAITECH DATASETS

Attention Module	UCSD Ped1	ShanghaiTech
Without	84.3%	69.4%
With	**84.9%**	**71.0%**

IV. CONCLUSION

We design an attention-based U-Net to build an encoder-decoder for anomaly detection. By inputting continuous video frames into the encoder-decoder, the next frame is predicted. The anomaly is detected by calculating the difference between the predicted frame and the real frame. The introduced attention module makes the model more focused on the foreground targets, and makes the normal foreground targets learn more fully during the training process. During testing, frequently occurring normal targets can be more accurately predicted , so that unusual targets that rarely appear are highlighted due to larger prediction errors. The experimental results on the public real surveillance videos show that the proposed method is a competitive anomaly detection model.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 61801417 and 61802336) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 18KJB520051).

REFERENCES

[1] M. Sabokrou, M. Khalooei, M. Fathy and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, 2018, pp. 3379-3388.

[2] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh and A. Van Den Hengel, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 2019, pp. 1705-1714.

[3] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," 2018 International Conference on Learning Representations(ICLR), Vancouver, 2018.

[4] H. Park, J. Noh and B. Ham, "Learning Memory-Guided Normality for Anomaly Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, 2020, pp. 14360-14369.

[5] W. Liu, W. Luo, D. Lian and S. Gao, "Future Frame Prediction for Anomaly Detection - A New Baseline," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, 2018, pp. 6536-6545.

[6] B. Ravi Kiran, Dilip Mathew Thomas, Ranjith Parakkal, "An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos," J. Imaging 4(2): 36 (2018).

[7] W. Luo, W. Liu and S. Gao, "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 341-349.

[8] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury and L. S. Davis, "Learning Temporal Regularity in Video Sequences," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 733-742.

[9] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," CoRR abs/1804.03999 (2018).

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, 2015, pp. 234-241.

[11] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, "Anomaly detection in crowded scenes," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR), San Francisco, CA, 2010, pp. 1975-1981.

[12] C. Lu, J. Shi and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," 2013 IEEE International Conference on Computer Vision(ICCV), Sydney, NSW, 2013, pp. 2720-2727.

[13] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," arXiv preprint arXiv:1511.05440, 2015.

[14] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," 2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Miami, FL, 2009, pp. 2921-2928.

[15] R. T. Ionescu, S. Smeureanu, B. Alexe and M. Popescu, "Unmasking the Abnormal Events in Video," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2914-2922.

[16] A. Dosovitskiy et al., "FlowNet: Learning Optical Flow with Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2758-2766.

[17] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, 2017, pp. 439-444.

[18] F. Wang et al., "Residual Attention Network for Image Classification," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6450-6458.