Enhanced Adversarial Learning Based Video Anomaly Detection with Object Confidence and Position

Yuxing Yang
Intelligent Sensing and Communications Research Group,
Newcastle University, UK
Email: y.yang60@newcastle.ac.uk

Zeyu Fu
Intelligent Sensing and Communications Research Group,
Newcastle University, UK
Email: z.fu2@newcastle.ac.uk

Syed Mohsen Naqvi
Intelligent Sensing and Communications Research Group,
Newcastle University, UK
Email: mohsen.naqvi@newcastle.ac.uk

***Abstract*—Video anomaly detection is to identify the abnormal objects, positions and behaviours during the video sequences. It is an important but challenging problem in intelligent video surveillance. Nowadays, there is much concern about the generative adversarial networks (GAN) to detect anomalies which contains two parts: generator and discriminator. However, the two networks of this model are hard to train well at the same time in practical use. In this paper, we propose to exploit object detection to enhance the adversarial learning model and to improve classification method to distinguish anomalies in a semi-supervised manner. We also detect object position anomaly in our proposed model which can not be done in generative adversarial learning models separately. The proposed framework is evaluated on dataset UCSD Ped1 and Ped2 using two criteria: area under the curve (AUC) and equal error rate (EER). The results confirm that our proposed method can effectively improve object variety anomaly performance and detect object position anomaly and is also superior to the baseline. Our approach also achieves improved performance compared with recent state-of-the-art methods.**

I. Introduction

Anomaly detection is important for modern intelligent video surveillance, with a huge impact in many fields such as public security, sports analysis and healthcare systems [1] [2] [3]. The rapidly increasing number of surveillance cameras make automated anomaly surveillance necessary since monitoring a huge number of cameras by operators may not be feasible but require a huge workload [4]. However, video anomaly detection is a challenging problem due to the infinite irregular objects and behaviours in the crowned scene [5]. Traditional video anomalies include abnormal objects such as bicycles and trucks in the sidewalk, abnormal locations such as pedestrians on the lawn and abnormal behaviours such as fighting or shooting [6].

One of the existing approaches for video anomaly detection is to use object-based detection [6] [7]. This method detects all regions and behaviors of interest objects in each frame and classifies whether the behavior of each object is normal or not. However, the limitation of this approach is that labeling training data and processing them are quite time-consuming and other existing problems can be false alarms and missed detections due to poor resolution of datasets. Another possible solution is frame-based classification [8] [9]. It focusses on whether the frame is normal or not but ignores the location and class of abnormal objects. Nevertheless, the method lacks understanding of the scenes so it may split a single object into pieces or mix different objects varieties into one patch [10]. As a typical adversarially learned one-class classifier (ALOCC) [8] proposed to generate fake data and observe patches likelihoods to discriminate the abnormality of the frame. This model can enhance the inlier objects and distort the outliers effectively. The training process of this approach needs to process a huge amount of video data to prevent the situation in which generator and discriminator are not trained well at the same time. Thus, this approach has a strict requirement for the dataset.

In this paper, we propose a novel video anomaly detection method by incorporating object detection information and patch likelihood results to discriminate whether the frame is abnormal or not. By exploiting object detection information with an adversarially learned anomaly detection framework, the discriminator can better recognize the anomalies. Test images split into patches to calculate the likelihood in the adversarial learning model. In addition, the detection model outputs the matrix of object varieties, bounding boxes and confidences. After an intersection over union (IoU) matching between patches and bounding boxes, each interset patch gets different weights according to confidence and a threshold is set to decide whether this frame is normal or not. Besides, the proposed approach can also discriminate abnormal location with the information of the object bounding box. The main contributions of this paper are threefold:

- A method based on adversarial learning and with the assistance of object detection is proposed for video anomaly detection.
- An improved classification method to discriminate where the video frame is normal or not.
- Abnormal objects positions are added to improve the performance of the algorithm.

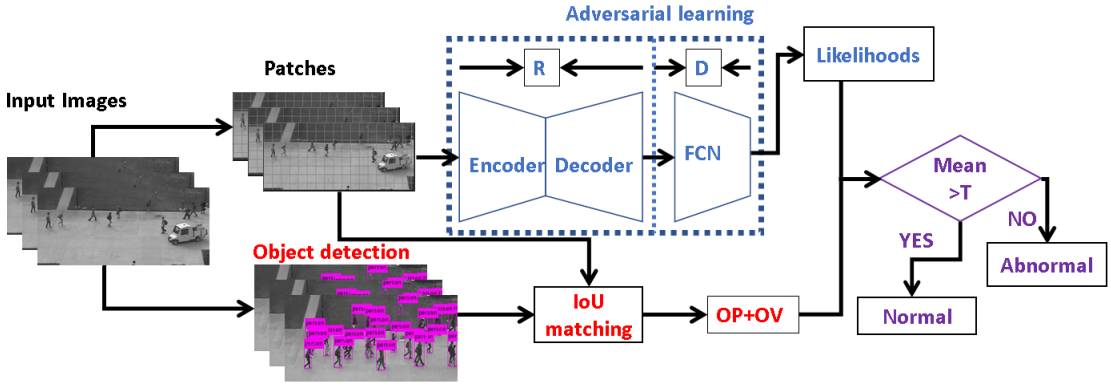

Fig. 1. Overview of the proposed method for video anomaly detection. There are two main streams in this framework. The adversarial learning model has two parts: the R network is to generate fake data and the D network is to discriminate the input data is real or fake. After training, the model can effectively discriminate anomaly varieties in each frame. The object detection model is to detect the objects information including varieties, bounding boxes and confidences. Then, these two models do a matching to discard background and only focus on the points of interest. Meantime, normal patches get high weight and abnormal varieties or positions patches get low level according to objects class, position and confidence results. Finally, we calculate the means of each frame and compare with the threshold to decide the frame is normal or not

The remainder of the paper is organized as follows: In Section 2, the proposed framework is introduced. Section 3 illustrates the setting of experiments and evaluate the performance of the proposed algorithm. Meanwhile, we compare it with the baseline method and discuss the results. Finally, Section 4 concludes the paper with future work.

II. The Proposed Anomaly Detection Algorithm

In this paper, we consider two types of possible anomalies, object-variety (OV) and object-position (OP) based anomalies. The object-variety method is based on ALOCC [8] model which has two parts: the R network and the D network. The R network is a generator and the D network is a discriminator. These two networks are trained adversarially in semi-supervised learning, in which the training data only contains normal video sequences. The object-position method is relying on the YOLOv3 [11] model to detect objects information including varieties, localizations and confidence scores. This approach can not only act as the prior information to help enhance the ALOCC model but also can output the frames of normal objects in abnormal positions.

A. Adversarial Learning Based Framework

ALOCC model is the basis of the whole anomaly detection framework. The basic theory of this model is the Generative Adversarial Networks (GAN) which contains two parts: generator and discriminator [12]. Generator network tries to generate fake samples that have the same distribution with real samples to fool discriminator and discriminator network is to distinguish real samples and fake samples. These two networks are learned adversarially and formulated as:

$$\binom{min}{G}\binom{max}{D}(E_{X\backsim p_t}[log(D(X))]+E_{Z\backsim p_Z}[log(1-D(G(Z)))]) \quad (1)$$

where Z is an input data, p_t is the real data distribution, Z is a random vector and p_Z is the input vector's distribution.

1) Training and Testing: Let I denote the input image and $\tilde{I}$ is the generated image. There are also two adversarial learning networks R and D which are similar to the GAN model. In the R network, $\bar{I}$ is generated instead of Z.

$$\tilde{I} = I + \eta \quad (2)$$

where η is the normal distribution noise with a mean 0 and standard deviation σ. Thus the η has distribution $N(0, \sigma^2 I)$ which is simplified as N_σ. This action can make R network robust to noise and distortion. Now the training criterion is updated as:

$$\binom{min}{R}\binom{max}{D}(E_{I\backsim p_t}[log(D(I))]+E_{\tilde{I}\backsim p_t+N_\sigma}[log(1-D(R(\tilde{I})))]) \quad (3)$$

To train the model, the model loss function L_{R+D} and reconstruction error L_R are optimized. The training step should stop at $L_R < p$ (p is a small positive number) which means R network can output fake data samples with the small difference compared with original ones. Since this is a semi-supervised learning method, there are only normal video frames in the training dataset, and the model must have the distribution P_t. In the testing phase, if the input patch is a normal object patch which has the similar distribution p_t as the model, after adding a normal distribution noise, $R(I)$ must approximately correspond to p_t because now the model's reconstruction error is very small. Thus, the output of discriminator of this kind of patches would be higher than discriminating original patches directly $D(R(I)) > D(I)$ which can be found in $a1$ and $a2$ in **Fig. 2(a)**. The R network's function is to denoise the input data. If the input patch is an outlier patch $\bar{I}$ and the distribution does not correspond to model's distribution, at this time, the reconstruction error would be large which means the R network can not reconstruct input patch well. Thus, the output of discriminator would be lower than discriminating normal input patches directly $D(R(\bar{I})) < D(R(I))$ which can be found from $a2$ and $b2$ in **Fig. 2(b)**.

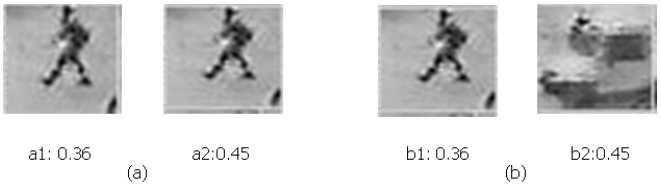

Fig. 2. Patches of original and generated data and correspond likelihoods. Figure(a) is the comparison of the patch likelihood results between direct classification method (a1) and adversarial learning classification method (a2). Figure(b) is the comparison of likelihood results between the normal patch (b1) and the anomaly (b2) under adversarial learning model.

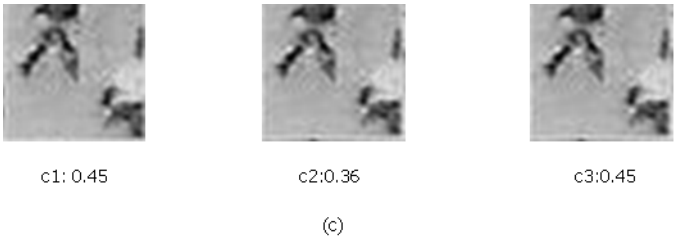

Fig. 3. The left figure (c1) is a mixed objects patch. It is hard for discriminator to distinguish it as shown in the middle figure. The likelihood of the middle figure (c2) is large than that of the left figure which is incorrect. After the proposed model, the likelihood of the right figure (c3) decreases as expected.

2) Justification about ALOCC Model: In summary, the above discussed ALOCC model can efficiently refine the likelihood of normal objects while it can also distort the likelihood of anomaly objects. Unfortunately, due to the lack of prior information of the frame, the procedure which splits frames to patches may segment one object into several patches, or a patch contains several objects such as $c1$ in **Fig. 3**, it would make the discriminator hard to decide the likelihood result. Besides, since the R network and the D network are trained at the same time, it is hard to ensure these two networks train well together. The common result is that the R network trains well while the D network does not get its best performance. Thus, some background patches which should be considered as normal patches have a low output likelihood. Besides, the evaluation of ALOCC model is in frame level which means if there is one pixel in a frame is abnormal, this frame is considered as an anomaly. Thus, an inferior discriminator affects the evaluation of the performance a lot.

B. Object Position and Variety Anomaly Detection (OP+OV)

To address the aforementioned problems, we propose a joint object detection and ALOCC framework which can detect novel object position anomaly and improve the ability of detecting object variety anomaly compared with the original ALOCC model. Object-position anomaly emphatically means the situation that a normal object in an abnormal position. This is relying on the YOLOv3 detection method which can detect all objects classes, bounding boxes and confidences in each frame [13]. The output matrix of detection is:

$$\mathbf{I}^k{}_j = [(C, x_1, x_2, y_1, y_2, S)_{1,j}, ..., (C, x_1, x_2, y_1, y_2, S)_{k,j}] \quad (4)$$

where k means the $k-th$ object and j means the $j-th$ frame and x_1 is left, x_2 is right, y_1 is top, y_2 is bottom of the bounding box,C is class and S is confidence score. By combining the information of the bounding box of each object, it is simple to decide whether the object is in an abnormal area or not by IoU calculation. IoU is a ratio of contact areas and total areas [3]. The objects bounding boxes information can match the patches in the test dataset, the object classes information can decide the positions of interest patches and the objects confidence information can remove false alarms from detection results. After those, only interest objects with high confidences remain to evaluate. The results of object variety and position anomaly detection are shown in **Fig. 4**.

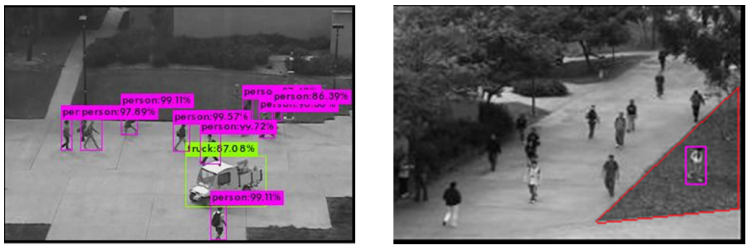

Fig. 4. Frames from Ped1 and Ped2 datasets. Frame(a) is object variety anomaly. The truck is the anomaly object. Frame(b) is object position anomaly. The frame is abnormal when the normal pedestrian is on an abnormal position such as lawn.

C. Improved Classification

Compared with only concerning the lowest discriminator likelihood of the frames in the ALOCC model [8], the classification approach we propose is more logical and robust. The specific steps are as follows: Firstly, all normal objects confidences should be compared with threshold α. This step can effectively reduce the influence of the detection of false alarms. Then, the IoU is calculated between S and each object in each frame which S is an advanced labeled abnormal area. This object can be discriminated as in an anomaly position if IoU value is over the threshold. We define this part of data as I'. This method is only suitable for a fixed camera dataset because it is hard to estimate which area is abnormal when the camera is mobile. The next, object bounding boxes should match the fixed patches of the test data. The contacted patches indexes should be recorded. This step can extract the desired object patches and reduce the influence of background. At last, according to confidence, normal and abnormal objects patches have different weights m and $1-m$, where m is the object confidence. And when the object is normal, the weights equals its confidence and when the object is abnormal, the weight equals $1-$confidence. If there are i normal patches, j abnormal variety patches and k abnormal position patches, the final formulation is:

$$P(I) = \frac{m(\sum_{i=0}^{i} D(R(I))) + (1-m) * (\sum_{j=0}^{j} D(R(\bar{I} + \sum_{k=0}^{k} D(R(I')))}{i+j+k} \quad (5)$$

where $P(I)$ is the average probability of the $I-th$ frame. Therefore, for testing frame I, the classification is as follows:

$$Frame(I) = \begin{cases} Normal & if\ P(I) > \tau \\ Abnormal & o.w. \end{cases} \quad (6)$$

III. Experiments

A. Dataset and Parameter Settings

We test our proposed framework on the UCSD benchmark dataset [14]. This dataset includes two subsets: Ped1 and Ped2. There are 34 training and 36 testing video sequences in Ped1, 16 training and 12 testing video sequences in Ped2. Since the dominant moving object in these two datasets are pedestrians, all non-pedestrian objects are considered as abnormal objects. Besides, pedestrians in the abnormal area such as lawn can also be considered as an anomaly. the output patches size of generative adversarial model is $45 * 45$ and the dimension of generative data vector is 100. The confidence threshold of the detection model is 0.2 to ensure all regions of interest are detected. And in the matching step, the normal patches confidence threshold is 0.6.

B. Evaluation Metrics

There are two evaluation criteria: area under the curve (AUC) and equal error rate (EER) [15]. Assuming test data only contains two classes: positive and negative. And the predicted results are only true or false. There are four classes: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The true positive rate and false positive rate can calculated as follows:true positive rate $(TPR) = TP/(TP + FN)$, false positive rate$(FPR) = FP/(TN + FP)$. The curve with TPR as y-axis and FPR as x-axis is Receiver Operating Characteristic (ROC) curve. The proportion under the ROC curve is AUC. And the intersection of ROC and auxiliary line $y = -x + 1$ is EER which means EER exists when $FPR = 1 - TPR$. Higher AUC and lower EER means better model performance.

TABLE I
EER(%) of Frame Level on Ped2 Dataset

Method	Ped2
SF [5]	42
MPPCA [16]	30
DanXu [17]	20
Conv-AE [18]	21.7
Cascade [19]	8.2
Proposed	12

C. Results and Discussions

Firstly, we conduct the performance analysis between the baseline and the proposed method. The performance of our algorithm in dataset Ped2 is better than the baseline [8]. From **Fig. 5**, the values of AUC have an obvious increase. And the EER values also decrease. Our test EER value approximately reaches 12%, a little less compared with the current best method Cascade [19]. As shown in **Table I**, the proposed method outperforms most of the state-of-the-art methods which confirms the advantages of the proposed method.

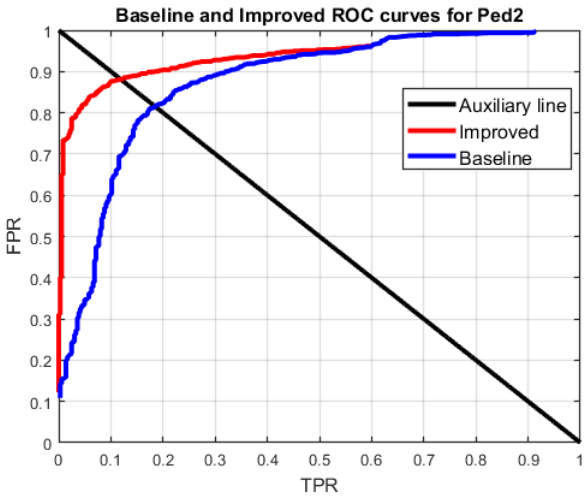

Fig. 5. ROC curves for Ped2 Dataset

Our proposed method can also provide a novel function that can detect object position anomaly. It is hard for a generative method to achieve this function since the generative model is not sensitive to object localization. The object position anomaly only occurs in Test9 and Test11 video sequences in Ped1 dataset. In **Fig. 6**, the blue curve shows a low performance for the basic generative model. It can find that the generative model is not efficient to object position anomaly. In our framework, with the aid of prior detection information, we can mitigate this problem. The red curve which is the performance of our proposed algorithm has great improvement compared with the blue curve of baseline model. However, the curves in **Fig. 6** is not smooth compared with the curves in **Fig. 5** since the anomaly position detection curves are tested on the two video sequences in Ped2 dataset, but the anomaly variety detection curves are tested in the whole Ped1 dataset. When we tested the whole Ped1 dataset, the EER and AUC have a low performance by the baseline and our proposed model. This is because the resolution of Ped1 is poorer than that of Ped2 and this dataset may not be suitable for the generative method for anomaly variety detection. Although these two video datasets are all taken using a fixed camera. During different viewpoints, the detection results and discriminator results all may take place significant changes. Therefore, how to improve the adversarial learning model and detection method are further work.

The problem of the baseline model is that the generator and discriminator are difficult to be well trained at the same time. Moreover, the situations that one patch contains both normal and anomaly objects are hard for the discriminator to take a decision. Thus, the prior detection method is necessary and helpful for the ALOCC model. From the results in **Fig. 5** and **Fig. 6**, the proposed method outperforms the baseline

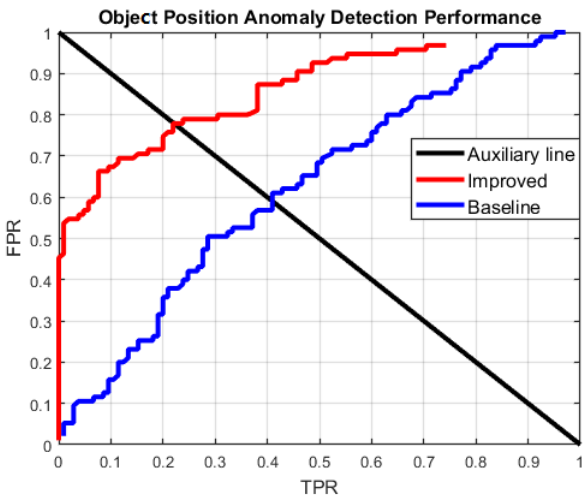

Fig. 6. ROC curves of OP Anomaly Detection for Test9 and Test11 Video Sequences in Ped1 Dataset

model in the datasets. There are three advantages of the proposed method. Firstly, the setting of confidence thresholds may reduce the effect of false alarms. Secondly, with the prior detection information, we can only extract the desired patches to reduce the background effect. The set of normal patches weights and abnormal patches weights would allow the discriminator to make decisions more easily. Finally, with the aid of frame detection, the normal objects in anomaly positions can be found which is hard in the baseline model.

IV. Conclusions

In this paper, we presented a novel method for video anomaly detection. With the aid of prior detection information, the model we proposed has a better performance to classify the pedestrians and other anomaly objects. The proposed method can also help to find the pedestrians anomaly positions such as crossing the lawn. Besides, we improve the evaluation method to distinguish whether the frame is normal or not. Detailed experiments were conducted which demonstrated the performance enhancement for GAN model training.

Observing the error data, most errors are caused by miss detection. This puts a high requirement on the quality of video datasets. Data argumentation method can be tried to deal with this limitation as future work.

References

[1] A. Kumbhar and P. C. Bhaskar, "Motion detection for video surveillance system," in *Advances in Computing and Data Sciences*, M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, and T. Ören, Eds. Singapore: Springer Singapore, 2018, pp. 523–534.

[2] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, "Particle phd filter based multiple human tracking using online group-structured dictionary learning," *IEEE Access*, vol. 6, pp. 14 764–14 778, 2018.

[3] Z. Fu, X. Lai, and S. M. Naqvi, "Enhanced detection reliability for human tracking based video analytics," in *International Conference on Information Fusion (FUSION)*, 2019, pp. 1–7.

[4] B. Antić and B. Ommer, "Video parsing for abnormality detection," in *International Conference on Computer Vision*, Nov 2011, pp. 2415–2422.

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

[6] F. Angelini, J. Yan, and S. M. Naqvi, "Privacy-preserving online human behaviour anomaly detection based on body movements and objects positions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 8444–8448.

[7] X. Li, W. Li, B. Liu, Q. Liu, and N. Yu, "Object-oriented anomaly detection in surveillance videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1907–1911.

[8] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] M. Sabokrou, M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli, "Avid: Adversarial visual irregularity detection," in *Computer Vision – ACCV 2018*, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham: Springer International Publishing, 2019, pp. 488–505.

[10] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, Sep. 2015.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[13] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[14] Z. Chen, Y. Tian, W. Zeng, and T. Huang, "Detecting abnormal behaviors in surveillance videos based on fuzzy clustering and multiple auto-encoders," in *IEEE International Conference on Multimedia and Expo (ICME)*, June 2015, pp. 1–6.

[15] 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, Jan 2014.

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

[17] D. Xu, R. Song, X. Wu, N. Li, W. Feng, and H. Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, pp. 144 – 152, 2014.

[18] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, April 2017.