Anomaly Event Detection Using Generative Adversarial Network for Surveillance Videos

Thittaporn Ganokratanaa*, Supavadee Aramvith*, and Nicu Sebe†
*Chulalongkorn University, Bangkok, Thailand
E-mail: supavadee.a@chula.ac.th Tel: +66-2 218 6909
†University of Trento, Trento, Italy
E-mail: niculae.sebe@unitn.it Tel: +39-0461 28 2989

***Abstract*— Anomalous event detection is advantageous for real-time video surveillance systems in terms of safety and security. Current works mostly run offline and struggle with abnormal event detection in crowded scenes. We propose unsupervised anomaly event detection using Generative Adversarial Network (GAN) with Optical Flow to obtain spatiotemporal features in appearance and motion representations. In training, GAN is used to train only the normal event images to generate their corresponding optical flow image. Hence, in testing, since the model knows only the normal patterns, any unknown events are considered as the anomaly event which can be detected by subtracting the pixels between the generated and the real optical flow images. We implement on the publicly available benchmark datasets and compare with state-of-the-art methods. Experiment results show that our model is effective for anomaly event detection in real-time surveillance videos.**

I. INTRODUCTION

Anomaly event detection in crowded scenes has increasingly gained interest rates in surveillance videos for public security in recent years [1, 2, 3, 4, 5, 6, 7]. Even though many abnormality detection works have been studied, there is still the unsolved issues. Anomaly event detection in crowded scenes is very challenging because of three main causes. The first one is about the small samples of anomaly dataset, causing the lack of information for data training in deep learning approach which needs a lot of data to gain high accuracy and to implement with various recognition tasks [8, 9, 10, 11]. The second issue is the absence of anomaly definition and objectives, while the third issue is about time complexity. Many abnormality works are suitable for offline use due to the high complexity system, making it incapable to run the system in real-time for the video surveillance system [12, 13, 14, 15, 16, 17, 18, 19, 20]. Moreover, these three issues are related to each other. The unclear of anomaly definition and objectives causes the difficulty of ground truth collection and implementation, resulting in higher costs.

To cope with these issues, generative approaches are attempted to detect abnormality by generating images based on the learning of normal events, which are what the model needs at training time. As the model knows no other patterns except the normal events, those other patterns are considered as the anomalous events. Thus, at testing time, the model is not able to generate abnormal patterns. Then, the anomalous event can be found by finding the difference from the learned patterns.

However, there is a lot of research on abnormality detection based generative approach counting on hand-crafted features [2, 3, 4, 7, 14], which leads to the difficulties in adapting to the real-world situations due to the limitation of user-defined features. M. Hasan et al. [15] proposed the autoencoder deep learning model with two networks. Firstly, hand-crafted features are extracted and then put into a fully connected autoencoder. Following that, the fully-connected autoencoder learns the features as a feed-forward neural network. Xu et al. [16] proposed an Appearance and Motion DeepNet (AMDN) for video abnormality detection by using stacked denoising autoencoders. However, their networks rely on small patches of images and need to train SVMs classifier for the learned model additionally.

In this paper, we propose the anomaly event detection based on the generative approach named Generative Adversarial Network (GAN) for the real-time video surveillance system. We aim to detect abnormality in crowded scenes. Instead of using GAN to generate new images, we use it to learn the spatiotemporal information of the normal patterns in crowds to obtain both appearance and motion features at the training time. Hence, the model is not able to generate the spatiotemporal representation of the unknown patterns, which are anomalous events, in the testing time. This training of normal pattern makes it easier for the model to detect any possible anomalous events. The anomalous events can be simply detected by subtracting the local pixels between the generated and the real testing images. Experimental results on the benchmark anomaly detection datasets show that our proposed method performs effectively for real-time anomaly event detection in crowds in terms of high accuracy and less time consumption compared to other state-of-the-art methods.

II. RELATED WORKS

Anomaly event detection research can be divided into two main approaches including a traditional and a deep learning-based method. The traditional-based method focuses on hand-crafted features [22, 23, 24, 25] which have limitations with complex scenes because of the difficulty of defining specific parameters for any possible abnormality patterns. Apart from

the traditional-based method, the deep learning-based method is widely used and more suitable for complex scenes. It has been studied in [20, 21] by using CNN to train recognition tasks. Ravanbakhsh et al. [20] proposed a final layer called a Binary Quantization Layer that plugs into the top of the network to gather motion information of anomaly patterns. Xu et al. [16] proposed appearance and motion feature representation by adapting autoencoders. However, the training is quite over-fitting and complicated due to the small anomaly datasets and the additional one-class SVMs classifier.

Our method is different from the methods mentioned above as we use only one network to focus on raw-pixels in the image in order to learn important features and train a generative network for the anomaly detection task based on the unsupervised deep learning method, which does not rely on hand-crafted features and any labeled samples. This unsupervised generative deep learning method learns the spatiotemporal features adaptively from the training data. Thus, it is more flexible and applicable to real-world use. We propose the spatiotemporal Generative Adversarial Network (GAN) for the detection of anomalous events in crowds. It is an effective approach that overcomes hand-crafted feature extraction and classification problems due to its outstanding performance [19] that can extract the significant features in the frames without any predefined anomaly types. In addition, GAN is a good approach for data augmentation and management because of its components, generator and discriminator networks, which help to prevent over-fitting and to train the deeper network on the end-to-end feature learning with small anomaly datasets.

In particular, GANs [26, 27, 28] consists of two networks, generator (G) and discriminator (D). G generates a new image (N) from an input image (I), and D tries to discriminate I from N. While G tries to fool D for producing more realistic image frames that are difficult to be discriminated. According to paper [29], the proposed framework provided the transformation from a sketch to a real-world image based on the use of conditional GANs with U-Net architecture. In contrast, we consider the transformation from the real-world to the motion pattern image which is not the realistic image. Hence, our G is used to learn the spatial pattern and transform it into the temporal pattern of the normal event, called spatiotemporal transformation. After the training, we can analyze the abnormal event from the generated temporal pattern image. However, even GAN outperforms other state-of-the-art methods, most of the literary works need to run offline to obtain good anomaly detection results. Thus, it is important to improve the computational cost of the abnormality detection system along with its accuracy for applying in real-time as it is a tradeoff between accuracy and time complexity. Our proposed method is specifically focused on improving the performance of abnormality detection for real-time surveillance videos.

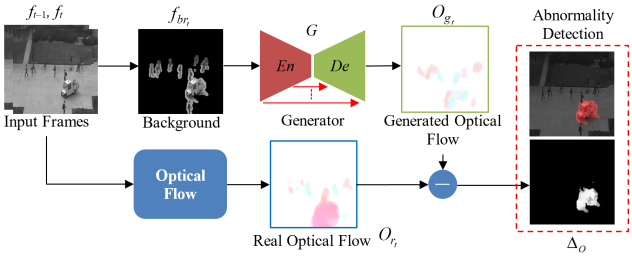


Fig. 1. Overview of proposed framework.

III. Spatiotemporal Translation Network for Normal Pattern Learning

We proposed a spatiotemporal translation GAN for abnormality detection in crowded scenes based on the image-to-image translation [29] which is able to model the mapping from the real image to the sketch image. In our proposed method, we map semantic spatial to temporal information by using the deep CNN of model G and D as shown in Fig. 1. On the other hand, the proposed GAN model is used to generate the optical flow (O_{g_t}) from the background removal frame (f_{br_t}) at time t, while the ground truth of optical flow (O_{r_t}) is obtained by using two consecutive frames in [30]. In addition, the G architecture comprises of Encoder (En) and Decoder (De) deep networks which are explained in detail in [29].

Intuitively, the input of G is an image x and a random noise z. However, this proposed framework uses the Dropout technique in the De of G to perform as the random noise z described in [29]. The output of G is a reconstructed image $g = G(x,z)$, which has the same dimension of the input image x but performs at a different channel. In detail, the input image of G is a background removal frame at time t ($x = f_{br_t}$) that is obtained by computing the frame absolute difference between two consecutive frames f_{t-1} and f_t. Then, G generates the output image g, which represents as the generated optical flow at the same time t ($g = O_{g_t}$), corresponding to the target image y ($y = O_{r_t}$). For the discriminator, D takes an input frame including the real optical flow O_{r_t} or the generated optical flow O_{g_t} to output a scalar signified the probability that the input frame derived from the real data.

At the training time, G and D are implemented on two objective functions; a Generator Loss L_{L1} and a GAN Loss L_{GAN}. G learns the mapping from x to y with the dropout noise z. Note that our network transforms only spatial to temporal data where the optical flow is defined by three-channel components including vertical direction, horizontal direction, and magnitude. The objective functions can be defined as below,

$$L_{L1}(G) = \mathrm{E}_{x,y,z}[\|y - G(x,z)\|_1], \quad (1)$$

$$L_{GAN}(G,D) = \mathrm{E}_y[\log D(y)] + \mathrm{E}_{x,z}[\log(1 - D(G(x,z)))]. \quad (2)$$

Finally, G is optimized as,

$$G^* = \arg\min_G \max_D L_{GAN}(G,D) + \lambda L_{L1}(G). \quad (3)$$

IV. Anomaly Detection

As G is used for learning normal events in the training, it is used for generating the output images in the testing with the same configuration parameters. At testing time, G is the only network used for reconstructing the learned features. It is input by each frame of the test video sequences, containing both normal and abnormal events. In this case, all the unknown events occurred in the scene are considered as the abnormality due to the fact that G knows only the normal events, resulting in the incapability of abnormality reconstruction. Following that, we can simply detect the abnormal patterns by computing the local difference in pixels between the frames of the real optical flow O_{r_t} and the generated optical flow O_{g_t}, represented as $\Delta_o = O_{r_t} - O_{g_t}$. The local difference Δ_o shows how much these two frames are different, assuring that G is unable to generate the abnormality patterns.

V. Experimental Results

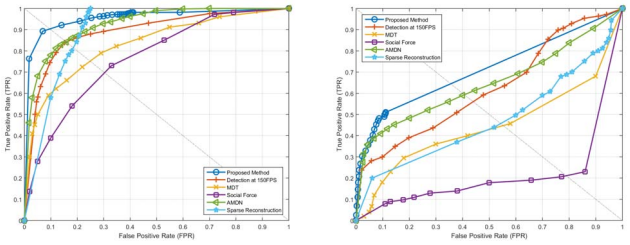
We discuss our implementation details, datasets, evaluation criteria, and experimental results including time complexity compared with other baseline methods. The experimental results are evaluated on two public anomaly datasets by using frame-level and pixel-level evaluations as same as the original parameter setting [1].

A. Implementation Details

At training time, we set the size of the frame from the training video of the UCSD dataset to 256 × 256 pixels. The reconstruction loss L_{L1} is optimized until 10^{-3} by using Adam optimization. We train our model on GPU NVIDIA GeForce GTX 1080 Ti, 484 GB/sec bandwidth with CUDA Cores 3584 and test on Intel Core i9-7960x CPU 2.8 GHz.

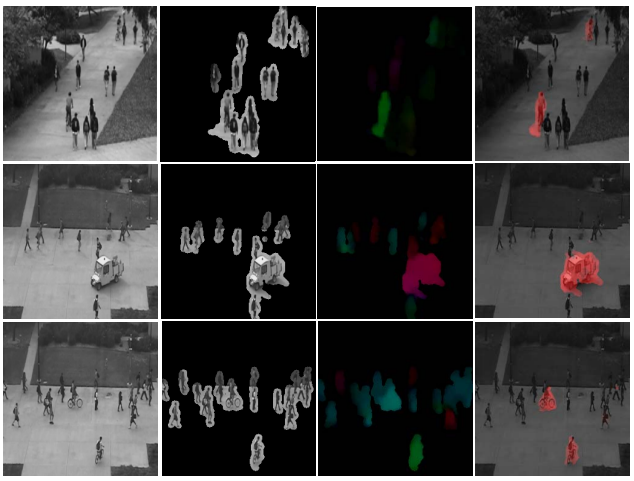
B. Datasets

There are two benchmark abnormality datasets used in this work including the UCSD dataset [3] and the UMN dataset [4]. The UCSD dataset is the realistic outdoor pedestrian scene in crowds, containing various abnormal events. It has two sub-folders: Ped1 and Ped2. Ped1 has 34 train and 16 test sequences, while Ped2 has 16 train and 12 test sequences. The UMN dataset contains the abnormality in crowds for both indoor and outdoor scenes. There are three different scenes in the total of 11 video sequences that contain 7,700 frames.



(a) Frame-level (b) Pixel-level

Fig. 2. ROC curves for both frame-level and pixel-level evaluations of UCSD Ped 1



(a) (b) (c) (d)

Fig. 3. Abnormal event detection on UCSD dataset: (a) original frame, (b) background removal frame, (c) optical flow frame, (d) abnormality detection

C. Evaluation Criteria

We evaluate the experimental results in both quantitative and qualitative results. The quantitative results of our method are evaluated by the frame-level and pixel-level evaluations and the qualitative results are visually represented in the standard protocol for anomaly detection as described below.

The frame-level evaluation checks that if at least one predicted anomalous pixel is in the frame, then the whole frame is labeled as the abnormal frame. The quantitative experimental results for frame-level of the UCSD Ped1 and Ped2 compared with other state-of-the-art methods using Area Under Curve (AUC) and Equal Error Rate (EER) are shown in Tab. 1. Receiver operating characteristic (ROC) curves of UCSD Ped1 are shown in Fig. 2. The comparison of quantitative results for the UMN dataset is shown in Tab. 2.

In contrast, the pixel-level evaluation focuses on the accuracy of anomaly position in the scene. This evaluation is challenging due to the complexity of locating abnormal areas. According to [1], the frame is a true positive when the true abnormality pixels are detected at least 40 percent over ground truth, otherwise it becomes a false positive.

TABLE I. EER AND AUC COMPARISON WITH STATE-OF-THE-ART METHODS ON UCSD DATASET

Method	Ped1 (Frame)		Ped1 (Pixel)		Ped2 (Frame)	
	EER	AUC	EER	AUC	EER	AUC
MPPCA [2]	40%	59.0%	81%	20.5%	30%	69.3%
SF+MPPCA [3]	32%	68.8%	71%	21.3%	36%	61.3%
MDT [3]	25%	81.8%	58%	44.1%	25%	82.9%
Social force (SF) [4]	31%	67.5%	79%	19.7%	42%	55.6%
Detection at 150 fps [5]	15%	91.8%	43%	63.8%	-	-
SR [7]	19%	-	54%	45.3%	-	-
AMDN (double fusion) [16]	16%	92.1%	40.1%	67.2%	17%	90.8%
Proposed Method	**9%**	**95.8%**	**35.6%**	**70.1%**	**9.8%**	**94.6%**

TABLE II. AUC COMPARISON WITH STATE-OF-THE-ART METHODS ON UMN DATASET

Method	AUC
Optical-flow [4]	0.84
SFM [4]	0.96
Sparse Reconstruction [7]	0.97
Commotion [31]	0.98
Plug-and-Play CNN [20]	0.98
Proposed Method	**0.99**

TABLE III. COMPUTATIONAL TIME COMPARISON WITH STATE-OF-THE-ART METHODS DURING TESTING (SECONDS PER FRAME)

Method	Ped 1	Ped 2
Sparse Reconstruction [7]	3.8	-
Commotion [31]	0.98	-
Detection at 150 fps [5]	**0.007**	-
IBC [32]	55	68
MDT [3]	17	23
Roshtkhari et al. [33]	0.16	0.18
Li et al. [34]	0.65	0.80
Xiao et al. [35]	0.22	0.29
STMC [36]	1.2	-
AMDN (double fusion) [16]	5.2	-
Proposed Method	0.086	**0.088**

The comparison of the quantitative results for the pixel-level evaluation of the UCSD Ped1 dataset with state-of-the-art methods is shown in Tab. 1. The ROC curves of the UCSD Ped1 dataset for the pixel-level evaluation are shown in Fig. 2. The quantitative results from Tab. 1 and Tab. 2 show that our proposed method achieves the best performance for anomaly event detection compared with other baseline methods.

Fig. 3 shows the qualitative results which are demonstrated as the images of anomaly event detection on the UCSD dataset. The detection areas of abnormal events are denoted as red pixels based on abnormality protocol visualization. From Fig. 3, it is clearly shown that our proposed method detects abnormal events in crowds effectively. The background removal helps to erase all the unnecessary objects (e.g., trees, roads) while retaining the important features of moving objects in the normal event (e.g., walking people). These features make *G* know more spatial information of the normal events, assisting in generating its corresponding temporal information. Since *G* is not able to reconstruct the anomalous objects and events (e.g., people riding a bike, driving a car, skateboarding), its generated optical flow is simply compared with the real optical flow to find the difference in local pixels.

In terms of the running time processing in frames per second (fps), we obtain 11.6 fps and 11.32 fps on the CPU for the UCSD Ped1 and Ped2, respectively. We compare our average running time processing in seconds per frame with other state-of-the-art methods as shown in Tab. 3. From Tab. 1, 2 and 3, our proposed method outperforms other state-of-the-art methods in both accuracy and running time aspects except for the running time in [5]. However, we achieve the lowest EER and the highest AUC compared with all state-of-the-art methods as shown in Tab. 1 and 2. According to our experimental results, our proposed method is suitable to apply with the real-time surveillance videos in crowds as it can effectively detect abnormal events with high-speed processing.

VI. CONCLUSIONS

In conclusion, we proposed an unsupervised spatiotemporal translation GAN network for real-time anomaly event detection in crowds. We aimed to increase the speed of anomaly detection and reduce system complexity. Our network is designed for learning both spatial and temporal features in its simplicity. As we train our network with only normal frames from the train video sequences of anomaly datasets, the network is not able to reconstruct the abnormality. So that at the testing time, the possible abnormal events are simply detected by computing the difference in pixels between the real and the generated motion frames. We implemented on two benchmark anomaly datasets. The experimental results show that our proposed method overcomes other state-of-the-art methods, considering both accuracy and time complexity. It can be applied in the realistic crowded scenes for real-time surveillance videos. For the future work, we will focus more on improving the accuracy of the anomaly event localization at the pixel-level evaluation.

ACKNOWLEDGMENT

This work is supported by Chulalongkorn University Dutsadi Phiphat Scholarship.

REFERENCES

[1] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *PAMI*, 2014.

[2] J. Kim and K. Grauman, “Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates,” in *CVPR*, 2009.
[3] V. Mahadevan, W. Li, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *CVPR*, 2010.
[4] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in *CVPR*, 2009.
[5] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in *ICCV*, 2013.
[6] V. Saligrama and Z. Chen, “Video anomaly detection based on local statistical aggregates,” in *CVPR*, 2012.
[7] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” in CVPR, 2011.
[8] R. B. Girshick, “Fast R-CNN,” in ICCV, 2015.
[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in ICML, 2014.
[10] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in CVPRW, 2014.
[11] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition,” in NIPS, 2014.
[12] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," Neurocomputing, vol. 219, pp. 548-556, 2017.
[13] Y. Yuan, Y. Feng, and X. Lu, "Structured dictionary learning for abnormal event detection in crowded scenes," Pattern Recognition, vol. 73, pp. 99-110, 2018.
[14] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, “Analyzing tracklets for the detection of abnormal crowd behavior,” in WACV, 2015.
[15] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 733-742: IEEE.
[16] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," Computer Vision and Image Understanding, vol. 156, pp. 117-127, 2017.
[17] S. Wang, E. Zhu, J. Yin, and F. Porikli, "Video anomaly detection and localization by local motion based joint video representation and OCELM," Neurocomputing, vol. 277, pp. 161-175, 2018.
[18] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," Neurocomputing, vol. 219, pp. 548-556, 2017.
[19] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in Image Processing (ICIP), 2017 IEEE International Conference on, 2017, pp. 1577-1581: IEEE.
[20] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1689-1698: IEEE.
[21] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, “Fully convolutional neural network for fast anomaly detection in crowded scenes,” arXiv:1609.00866, 2016.
[22] A. Li, Z. Miao, Y. Cen, and Q. Liang, "Abnormal event detection based on sparse reconstruction in crowded scenes," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, 2016, pp. 1786-1790: IEEE.
[23] R. V. H. M. Colque, C. A. C. Júnior and W. R. Schwartz, "Histograms of Optical Flow Orientation and Magnitude to Detect Anomalous Events in Videos," 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, 2015, pp. 126-133.
[24] K. Cheng, Y. Chen and W. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2909-2917.
[25] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares and F. Brémond, "Toward Abnormal Trajectory and Event Detection in Video Surveillance," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 3, pp. 683-695, March 2017.
[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in NIPS, 2014.
[27] Tim Salimans, I. J. Goodfellow, Wo. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in NIPS, 2016.
[28] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” ICLR, 2016.
[29] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *CVPR*, 2017.
[30] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool, "Fast optical flow using dense inverse search," in European Conference on Computer Vision, 2016, pp. 471-488: Springer.
[31] H. Mousavi, M. Nabi, H. Kiani, A. Perina, and V. Murino, “Crowd motion monitoring using tracklet based commotion measure,” in ICIP, 2015.
[32] O. Boiman and M. Irani, “Detecting irregularities in images and in video,” Int. J. Comput. Vis., vol. 74, pp. 17–31, Aug. 2007.
[33] M. J. Roshtkhari and M. D. Levine, “An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions,” Comput. Vis. Image Understand., vol. 117, no. 10, pp. 1436–1452, Oct. 2013.
[34] W. Li, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 1, pp. 18–32, Jan. 2014.
[35] T. Xiao, C. Zhang, and H. Zha, “Learning to detect anomalies in surveillance video,” IEEE Signal Process. Lett., vol. 22, no. 9, pp. 1477–1481, Sep. 2015.
[36] Y. Cong, J. Yuan, and Y. Tang, “Video anomaly search in crowded scenes via spatio-temporal motion context,” IEEE Trans. Inf. Forensics Security, vol. 8, no. 10, pp. 1590–1599, Oct. 2013.