Anomaly Detection in Surveillance Videos by Future Appearance-motion Prediction

Tuan-Hung Vu[1], Sebastien Ambellouis[2], Jacques Boonaert[1] and Abdelmalik Taleb-Ahmed[3]
[1] *Departement Informatique & Automatique, IMT Lille Douai, France*
[2] *COSYS, IFSTTAR, France*
[3] *IEMN DOAE UMR CNRS 8520, Université Polytechnique Hauts-de-France, France*

Keywords: Anomaly Detection, Future Prediction, Deep Learning, Appearance and Motion Features.

Abstract: Anomaly detection in surveillance videos is the identification of rare events which produce different features from normal events. In this paper, we present a survey about the progress of anomaly detection techniques and introduce our proposed framework to tackle this very challenging objective. Our approach is based on the more recent state-of-the-art techniques and casts anomalous events as unexpected events in future frames. Our framework is so flexible that you can replace almost important modules by existing state-of-the-art methods. The most popular solutions only use future predicted informations as constraints for training a convolutional encode-decode network to reconstruct frames and take the score of the difference between both original and reconstructed information. We propose a fully future prediction based framework that directly defines the feature as the difference between both future predictions and ground truth informations. This feature can be fed into various types of learning model to assign anomaly label. We present our experimental plan and argue that our framework's performance will be competitive with state-of-the art scores by presenting early promising results in feature extraction.

1 INTRODUCTION

Automatic anomaly detection in video sequence is a very important task for smart security systems, especially in transportation or security application fields. This work recently has became active research topic because of the necessity of automatic anomaly detection in real-world context. Actually, the frequency of abnormal events is really rare compared with normal events and its features usually do not follow any spatial or temporal relation. Thus, we need a huge resources, not only the workers but also time-consuming to manually process the anomaly detection task. Therefore, our work is significant in term of reducing processing cost for real-world systems.

Naturally, in human behavior analysis, we might consider anomaly detection as an action recognition problem. But this classical point of view lead us to an unbalanced situation where the number of samples for each class is significantly different. Beside, it is difficult to pre-define the structure of abnormal events because there is usually not any spatial and temporal relations between those events. Hence, we should tackle this challenge in a specific way. Generally, from the first successful works until now, they proposed three solutions: one-class classification based (Wang and Snoussi, 2012; Wang and Cherian, 2019), changing detection based (Giorno et al., 2016; Hasan et al., 2016; Ionescu et al., 2017) and future prediction based (Nguyen and Meunier, 2019; Liu et al., 2017).

One-class learning first constructs the representation for events then fit a model to data for which annotations are available only for a single class, normally those are labels for abnormal samples. This solution is only appropriated for binary classification and it has limitation when we need further information as type and localization. Changing detection is a classical way where each event is compared with its neighbors to find the most different ones. By this way, we could get trouble when abnormal event always or never happens in a sequence. The future prediction based techniques casts abnormal events as unpredicted events. A generative model to produce future information from previous frames is computed and a model is trained from normal frames and noisy ones; usually more

Vu, T., Ambellouis, S., Boonaert, J. and Taleb-Ahmed, A.
Anomaly Detection in Surveillance Videos by Future Appearance-motion Prediction.
DOI: 10.5220/0009146704840490
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 484-490
ISBN: 978-989-758-402-2; ISSN: 2184-4321

noisy frames are more blurred than the ground truth (Figure 1).

Recently, thanks to the powerful performance of deep learning models as auto encode-decode joint to generative adversarial learning, future prediction based methods achieved state-of-the-art for many anomaly challenge: CUHK avenue (Lu et al., 2013), UCSD pedestrian (Mahadevan et al., 2010), ShanghaiTech (Liu et al., 2017), etc. In these methods future predicted information are only used as constraints for training a convolutional encode-decode network. Moreover, the abnormal classification decision is computed by thresholding the score of the difference between original and the reconstructed information at the current frame. Inspired by this promising solution, we proposed a fully future prediction based framework that directly consider the difference between future predictions and ground truth informations as features. After histogram encoding, these representations feed into various types of learning model to assign the anomaly labels. The rest of the paper is presenting the following contributions:

- A short survey about anomaly detection and future prediction methods
- A presentation of our fully future prediction based and flexible framework for anomaly detection
- A definition of our feature vector used for anomaly detection: histogram of future appearance-motion difference (HOFAMD)
- An introduction to some learning techniques based on HOFAMD

This paper will be ended by a presentation of the evaluation plan and some conclusions and perspectives.

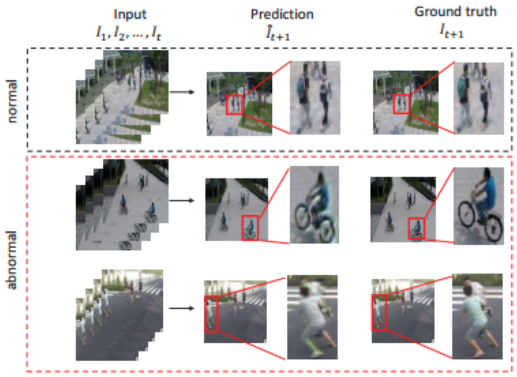

Figure 1: Some predicted frames and their ground truth in normal and abnormal events (Liu et al., 2017).

2 RELATED WORK

In this section, we present a survey about anomaly detection methods and future information prediction techniques on which the most promising solutions and our framework solutions are based.

2.1 Anomaly Detection

Anomaly detection methods can be splitted into two categories depending on the type of the features used to characterize abnormal and normal events:

- Hand-crafted features (Kim and Grauman, 2009; Mahadevan et al., 2010; Giorno et al., 2016; Wang and Snoussi, 2012; Medioni et al., 2001; Zhang et al., 2009)
- Deep learning features (Hasan et al., 2016; Hinami et al., 2017; Luo et al., 2017; Ionescu et al., 2017; Liu et al., 2017; Nguyen and Meunier, 2019).

On the one hand, before the rise of Convolutional Neural Networks (CNNs), most of the methods was extracting hand-crafted features to finally estimate the clusters of normal and abnormal events distributions. In some early works, the principal features is motion trajectories (Medioni et al., 2001; Zhang et al., 2009) and it executes quite fast and simple for implementation. But its performance always depend on the quality of the detectors and the trackers which are easily confused in crowed and complex scenes. Moreover, the only coordinates is not sufficient to describe all the spectrum of abnormal events. To deal with this problem, information about appearance and motion are extracted along the trajectories. Histogram of optical flow was used by (Kim and Grauman, 2009) to build space-time Markov Random Fields graph. (Mahadevan et al., 2010) learned the Mixture of Dynamic Textures (MDT) during training then computed negative log-likelihood of the spatio-temporal patch at each region at test phase. (Wang and Snoussi, 2012) built Histograms of optical flow orientation (HOFO) then classified events by one-class SVM or kernel PCA. A combination of HOG, HOF, MBH was used by (Giorno et al., 2016) to train their classifiers then take the average classification scores to draw the output signal.

On the other hand, the progress of deep learning method lead to many successful researches in anomaly detection. (Hasan et al., 2016) utilized either motion trajectories features (HOG, HOF, MBH) or learned features combined with autoencoder to reconstruct the scene. The reconstruction error is used to measure the regularity score that can be further analyzed for different applications. (Hinami et al., 2017) integrated a generic Fast R-CNN model and environment-dependent anomaly detectors. The authors first learn CNN with multiple visual tasks to exploit semantic information that is useful for detecting and recounting abnormal events then appropriately plugged the model into anomaly detectors. (Ionescu et al., 2017) combines the motion features computed from 3D gradients at each spatio-temporal cube with *conv5* layer of VGG-net with fine-tuning as appearance features. Then a binary classifier is trained to distinguish between two consecutive video sequences while removing at each step the most discriminant features. Higher training accuracy rates of the intermediately obtained classifiers represented abnormal events. (Luo et al., 2017) proposes a Temporally-coherent Sparse Coding (TSC) where they enforce similar neighboring frames be encoded with similar reconstruction coefficients. Then the authors mapped the TSC with a special type of stacked Recurrent Neural Network (sRNN). (Liu et al., 2017) introduces a first work of future prediction based anomaly detection. They adopted U-Net as generator to predict next frame. To generate high quality image, they made the constraints in terms of appearance (intensity loss and gradient loss) and motion (optical flow loss). Then the difference between a predicted future frame and its ground truth is used to detect an abnormal event. In (Nguyen and Meunier, 2019), the authors continue this approach by designing a model as a combination of a reconstruction network and an image translation model that share the same encoder. The former sub-network determines the most significant structures that appear in video frames and the latter one attempted to associate motion templates to such structures. They achieve state-of-the-art for 6 popular benchmarks of anomaly detection.

2.2 Future Prediction

Future video informations prediction recently has became an active topic due to significant progress in deep learning, especially in generative adversarial networks (GANs) and Convolutional Auto-Encode (Conv-AE) models. They predicted various type of future informations for specific applications. (Mathieu et al., 2015) trained a classical 7-layers CNN to generate future frames given an input sequence. To deal with the inherently blurred predictions obtained from the standard Mean Squared Error (MSE) loss function, they proposed three different and complementary feature learning strategies: a multi-scale architecture, an adversarial training method, and an image gradient difference loss function. (Walker et al., 2015) built a 7-layers CNN for predicting the future motion of each and every pixel in the image in terms of optical flow given a static image. (Finn et al., 2016) developed a Long-Short Term Memory (LTSM) based action-conditioned video prediction model that explicitly models pixel motion to learn about physical object motion without labels, by predicting a distribution over pixel motion from previous frames. Inspired by the same idea, (Lotter et al., 2016) constructed LTSM based PredNet which learned to predict future frames in a video sequence, with each layer in the network making local predictions and only forwarding deviations from those predictions to subsequent network layers. (Villegas et al., 2017) built a deep neural network for the prediction of future frames in natural video sequences upon the Conv-AE and Convolutional LSTM for pixel-level prediction, which independently capture the spatial layout of an image and the corresponding temporal dynamics. In (Oliu et al., 2017), authors introduced an architecture based on recurrent Conv-AEs to deal with the network capacity and error propagation problems for future video prediction. It consisted on a series of bijective Gate Recurrent Unit (GRU) layers, which allowed for a bidirectional flow of information between input and output: they considered the input as a recurrent state and update it using an extra set of gates. (Gao et al., 2017) proposed an approach using Conv-AE that hallucinated the unobserved future motion implied by a single snapshot to help static-image action recognition. The key idea was to learn a prior over short-term dynamics from thousands of unlabeled videos, infer the anticipated optical flow on novel static images, and then train discriminative models that exploit both streams of information. Obviously, most of recent researches build their model upon a Conv-AE model to reconstruct the future informations.

3 PROPOSED METHODS

In this section, we describe our framework for anomaly detection in detail. The general pipeline is shown on Figure 2. This pipeline is so flexible that we can replace any components (appearance reconstructor, optical flow estimator-generator, learning model) by the state-of-the-art methods, the one that will be considered as the more adapted to the application context.

3.1 Future Appearance Reconstruction

We build a Conv-AE using U-Net structure and follow the successful model of (Liu et al., 2017). Instead of

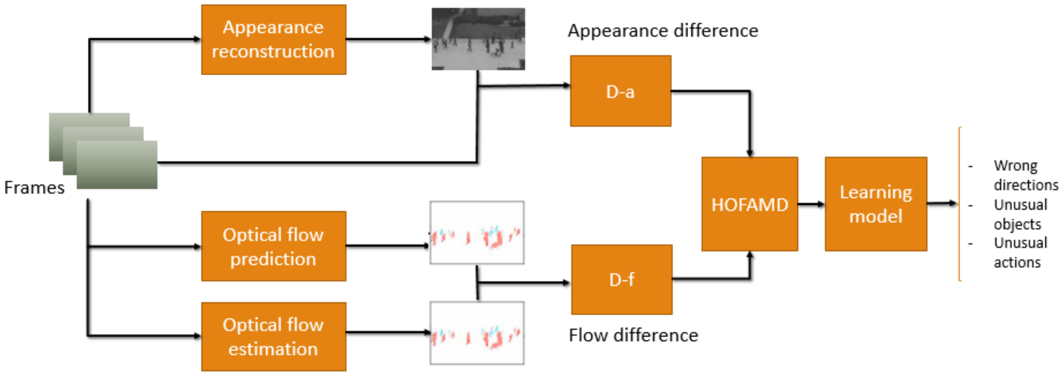

Figure 2: General pipeline of our anomaly detection framework. A sequence of frames feed into 3 streams: Appearance reconstruction stream is an Conv-AE model to predict future RGB frame from previous frames; Optical flow estimation stream is a strong flow estimator such as HD3 (Yin et al., 2019) to produce ground truth flow; Optical flow prediction stream is a strong flow predictor such as Im2Flow (Gao et al., 2017). The output of first stream is compared with original frame to calculate appearance difference. The outputs of the two other streams are compared together to calculate motion difference. All those differences parameters are encoded to histogram then fused into Histogram of Future Appearance-Motion Difference (HOFAMD) descriptor. This representation thus will be used for some learning techniques for anomaly detection. Some small images are taken from (Nguyen and Meunier, 2019).

using the difference between the output and the original versions of video frame as well as the optical flow together for adversarial training, we simple the net by directly taking into account the appearance difference of the reconstructed I^r and original frame I as constraint. We also keep the same idea of taking the sum of intensity loss and gradient loss along both x, y dimension as appearance loss.

$$L_{int}(I^r, I) = \| I^r - I \|_2^2 \tag{1}$$

$$L_{grad} = \sum_{x,y} \| \, |grad_{x,y}(I^r)| - |grad_{x,y}(I)| \, \| \tag{2}$$

$$L_{appr} = L_{int} + L_{grad} \tag{3}$$

The network configuration of this component is the same as appearance encoder-decoder of (Nguyen and Meunier, 2019). The appearance reconstructor is trained with video frames of normal events so it will produce the output appropriated to normal events. The idea is that the abnormal frames would not be well-reconstructed so the higher difference between the reconstruction and ground truth is produced.

3.2 Future Motion Prediction

The motion predictor is taken as Im2Flow (Gao et al., 2017). This framework achieved state-of-the-art for flow prediction. In detail, the network structure of Im2Flow is similar to the optical flow encode-decode branch of (Nguyen and Meunier, 2019). But instead of only using l_1 loss between predicted flow and ground truth flow, Im2Flow considered the combination of two losses: a pixel error loss and a motion content loss. The pixel loss measured the agreement with the true flow while the motion content loss enforced that the predicted motion image preserved high level motion features. This improvement might help Im2Flow worked better. The ground truth flow is computed by HD3 (Yin et al., 2019), top 10 methods for optical flow estimation on KITTI flow benchmark.

3.3 Descriptor Encoding

We consider both appearance and motion difference for feature encoding. Given a sliding window M with size $W \times H$, we first compute the square distance at pixel level between both reconstructed appearance I^r and predicted flow F^p with ground truth appearance I and flow F.

$$D_a = \frac{1}{W \times H} \sum_{i,j \in M} (I^r_{i,j} - I_{i,j})^2 \tag{4}$$

$$D_f = \frac{1}{W \times H} \sum_{i,j \in M} (F^p_{i,j} - F_{i,j})^2 \tag{5}$$

Then we normalize D_a and D_f with the maximum value for each type over a frame.

$$|D_a| = \frac{D_a}{\arg\max_I D_a} \tag{6}$$

$$|D_f| = \frac{D_f}{\arg\max_F D_f} \tag{7}$$

We divide the range $[0,1]$ by N bins then we use a voting strategy to histogram encode for both $|D_a|$ and $|D_f|$. We have 3 channels RGB for D_a and 3 channels *x,y*, magnitude for D_f, so the size of D_a and D_f is $3 \times N$. Then we combine them with the weights $\leftthreetimes_a$ and $\leftthreetimes_d$. Each weight is the inverse of the average of D_{max} over a n-frame sequence.

$$\leftthreetimes = (\frac{\sum D_{max}}{n})^{-1} \tag{8}$$

We propose two ways to combine them. The first way just concatenates them as $[\leftthreetimes_a D_a, \leftthreetimes_f D_f]$ to obtain a $2 \times 3 \times N$ dimension vector. In the second way, we take the weighted sum as $[\leftthreetimes_a D_a + \leftthreetimes_f D_f]$ and the dimension of the vector is $3 \times N$. We call this last feature combination as Histogram of Future Appearance-Motion Difference (HOFAMD).

3.4 Learning Models for Anomaly Detection

As discuss above, HOFAMD can feed into various learning techniques. We present 3 highlight methods: learn a simple threshold, changing detection and clustering based one-vs-rest SVM.

By the first method, we simply our feature descriptor into an anomaly scores then learn an appropriated threshold for detection as strategy of (Nguyen and Meunier, 2019). If we set $N = 1$ then we take the sum of all 3-channels for only the maximum value, we have one final score for each frame. This score can be easily compared with a threshold and the higher value determines abnormal events.

By the second method called changing detection, each sequence is compared with its neighbors by applying strategy of (Ionescu et al., 2017). We iteratively train a binary classifier to discriminate between two consecutive video sequences while discarding at each step the most discriminant features. Higher training accuracy rates of the intermediately obtained classifiers represent abnormal events.

The last learning technique has been recently proposed by (Ionescu et al., 2018). We follow the key idea of considering the problem as supervised learning where they did the clustering on the training samples into normality clusters. Then, a one-versus-rest abnormal event classifier was employed to discriminate each normality cluster from the rest. For the objective of training the classifier, the other clusters acted as dummy anomalies.

4 EVALUATION

In this section, we present the 3 popular benchmarks for anomaly detection on which we will evaluate our work. We finally describe our experiment setup plan.

4.1 Dataset

We are evaluating our framework on 3 popular benchmarks: CUHK avenue (Lu et al., 2013), UCSD pedestrian (Mahadevan et al., 2010) and ShanghaiTech (Liu et al., 2017).

- CUHK avenue has 16 training and 21 testing videos containing 47 irregular events, including throwing objects, loitering and running. The size of people may change because of the camera position and angle.
- UCSD pedestrian includes Ped1 and Ped2. Ped1 has 34 training and 36 testing videos with 40 abnormal events. All of these anomalous cases are about vehicles such as bicycles and cars. Ped2 has 16 training and 12 testing videos with 12 abnormal events. The definition of anomaly for Ped2 is the same with Ped1.
- Shanghai Tech contains 13 scenes integrated complex light conditions and camera angles. It includes 130 abnormal events and over 270, 000 training frames. Moreover, pixel level ground truth of abnormal events is also annotated.

4.2 Experiment Plan

Our framework will be largely evaluated on all 3 benchmark. The size of sliding window is set to 16×16 as proposed in (Nguyen and Meunier, 2019). The N-number of bins for histogram encoding is also evaluated for $N = 1,4,8$. Both combining strategies of HOFAMD are tested and report the performance. To find the most appropriated learning techniques, we will apply HOFAMD for all 3 methods above and analysis the performance. The Area Under Curve (AUC) is utilized as evaluation metric.

4.3 Early Results of Feature Extraction

We use HD3 (Yin et al., 2019) as optical flow estimator to extract flow ground truth from sequences of CUHK Avenue dataset. Then we implement Im2Flow (Gao et al., 2017) as optical flow predictor to predict future flow from the same sequences. Results of a sequence containing abnormal action are illustrated in Figure 3. The abnormal action is ”a man is running fast from right side to left side”. We find

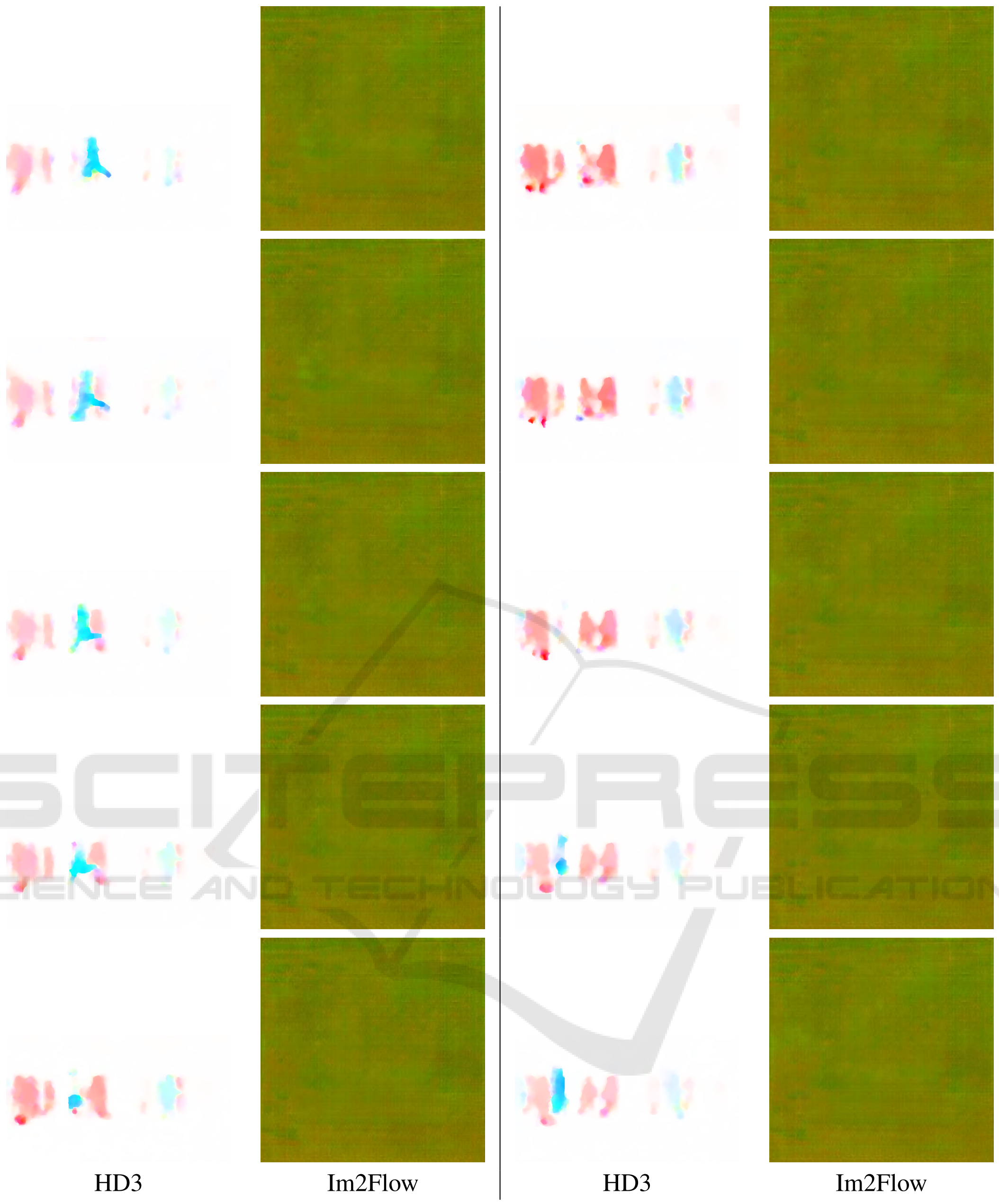

Figure 3: Results of optical flow estimation and prediction performed on a sequence of CUHK avenue dataset containing abnormal action.

that the normal action (people walking) is estimated almost well while abnormal action (person running) is blurred in some images due to the large displacement. In contract, both types of action are neutralized by optical flow predictor. Hence, when we take the subtraction of estimation flow and prediction flow, the difference between abnormal and normal action will be highlighted. It means that the $D-f$ block can be a promising representation for anomaly detection.

5 CONCLUSIONS

In conclusion, we first do a survey about progress of anomaly detection then present a flexible future prediction based framework whose components benefited from state-of-the-art methods. We also propose a histogram based feature called HOFAMD which represents for the difference of predicted information and ground truth including appearance and motion. To evaluate the performance of HOFAMD, we introduce

3 useful learning techniques for anomaly detection.

For future work, the very next steps is evaluate our framework on 3 benchmarks following experiment plan. Then we investigate to integrate as much as components for an end-to-end network.

REFERENCES

Finn, C., Goodfellow, I. J., and Levine, S. (2016). Unsupervised learning for physical interaction through video prediction. *NIPS*.

Gao, R., Xiong, B., and Grauman, K. (2017). Im2flow: Motion hallucination from static images for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5947.

Giorno, A. D., Bagnell, J. A., and Hebert, M. (2016). A discriminative framework for anomaly detection in large videos. In *ECCV*.

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

Hinami, R., Mei, T., and Satoh, S. (2017). Joint detection and recounting of abnormal events by learning deep generic knowledge. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3639–3647.

Ionescu, R. T., Khan, F. S., Georgescu, M.-I., and Shao, L. (2018). Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *CVPR*.

Ionescu, R. T., Smeureanu, S., Alexe, B., and Popescu, M. (2017). Unmasking the abnormal events in video. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2922.

Kim, J. and Grauman, K. (2009). Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928.

Liu, W., Luo, W., Lian, D., and Gao, S. (2017). Future frame prediction for anomaly detection - a new baseline. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6536–6545.

Lotter, W., Kreiman, G., and Cox, D. D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *ICLR*.

Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab.

Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked rnn framework. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981.

Mathieu, M., Couprie, C., and LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *ICLR*, abs/1511.05440.

Medioni, G., Cohen, I., Bremond, F., Hongeng, S., and Nevatia, R. (2001). Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889.

Nguyen, T. N. and Meunier, J. (2019). Anomaly detection in video sequence with appearance-motion correspondence. *ICCV*.

Oliu, M., Selva, J., and Escalera, S. (2017). Folded recurrent neural networks for future video prediction. In *ECCV*.

Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. (2017). Decomposing motion and content for natural video sequence prediction. *ICLR*, abs/1706.08033.

Walker, J., Gupta, A., and Hebert, M. (2015). Dense optical flow prediction from a static image. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2443–2451.

Wang, J. and Cherian, A. (2019). Gods: Generalized one-class discriminative subspaces for anomaly detection. *ICCV*.

Wang, T. and Snoussi, H. (2012). Histograms of optical flow orientation for visual abnormal events detection. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 13–18.

Yin, Z., Darrell, T., and Yu, F. (2019). Hierarchical discrete distribution decomposition for match density estimation. In *CVPR*.

Zhang, T., Lu, H., and Li, S. Z. (2009). Learning semantic scene models by object classification and trajectory clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1947.