Video Anomaly Detection for Traffic Accidents Based on Frame Prediction

Guangyu Sun
University of Rochester
gsun6@ur.rochester.edu

Yangxin Fan
University of Rochester
yfan24@ur.rochester.edu

Abstract

Road traffic control has existed for a long time to safeguard vehicles and pedestrians. Detecting anomalies in traffic videos can help us to make our road traffic control systems safer and act swiftly to save lives and properties when fatal traffic accidents occur. However, there is no efficient system accurate enough to detect traffic accidents and help drivers or pedestrians to avoid such tragedies. To solve this issue, the following study proposes to use an existing future frame prediction based method, which has already been successfully applied to solve anomalies detection for pedestrians in sidewalks [12], to detect and predict accidents in live traffic videos. Different from most traffic anomalies detection methods that are based on minimizing reconstruction errors, the future frame based method takes both spatial and motion constrains into consideration, which fits better to the essence of video itself with both sudden appearance and quick motion being likely anomalies. Extensive experiments on our selected CADP dataset [19] validate the effectiveness of our newly introduced method in video anomaly detection for traffic accidents. Our model is trained by normal traffic frames and can detect and predict traffic accidents.

1. Introduction

Intelligent traffic system with surveillance cameras for detecting traffic accidents can be an important component of smart city or AI City development. Accurately detecting traffic accidents through running computer vision algorithms on live traffic videos has huge advantages over hiring human labors to monitor surveillance cameras system to detect anomalies. First, it can help police to locate and investigate traffic accidents more quickly. Second, it reduces human errors especially when human labors get exhausted in tedious monitoring work. Third, if fatal accidents happen, the ambulances and medical personnel from local hospitals could be sent more quickly to save lives. Last but not the least, efficient traffic accidents detecting can help self-driving system like autopilot to predict traffic accidents beforehand.

Anomaly detection is an important task to judge the videos efficiently. An ordinary approach for traffic anomalies detection is to detect stopped/static vehicles. However, such methods show a gap in anomaly detection as they detect an abnormal event when a stalled vehicle already stops. Besides, the algorithm only captures stalled vehicles after a certain timespan, which leads to a significant disadvantage since a large number of static vehicles can be resulted from accidents. Another shortcoming for normal method is due to factors like viewpoints, weather and lighting conditions, and etc, affecting qualities of real traffic videos. The current start-of-the-art methods still need expensive training resources and often fail to predict future anomalies due to huge imbalance between normal and anomalous video frames in real scenarios. Traffic surveillance cameras usually have fixed angles so that they will capture similar scenes, which makes this task possible to solve.

We summarize our contribution as follows: i): We applied the future frame prediction [12] based method for traffic video anomaly detection to detect traffic accidents. This method imposes both spatial constraints (intensity and gradient loss) and temporal constraints (optical flow loss) in the training loss function to capture the essence of two kinds of abnormal events: appearance and motion, see Figure. 1. ii): We improved future frame based method's prediction accuracy through improving calculation method of anomaly scores.

The dataset we use is Car Accident Detection and Prediction (CADP) [19], which consists of 1,416 video segments collected from YouTube, with 205 video segments have full spatio-temporal annotations. It is a novel dataset for traffic accidents analysis, which aims to resolve the lack of public data for research about automatic spatio-temporal annotations for traffic safety in the roads.

2. Related Work

Due to limited availability of well labeled and sufficient anomalous instances, unsupervised or semi-supervised methods are commonly adopted [10]. They learn structures of normality through nominal training videos and attempts

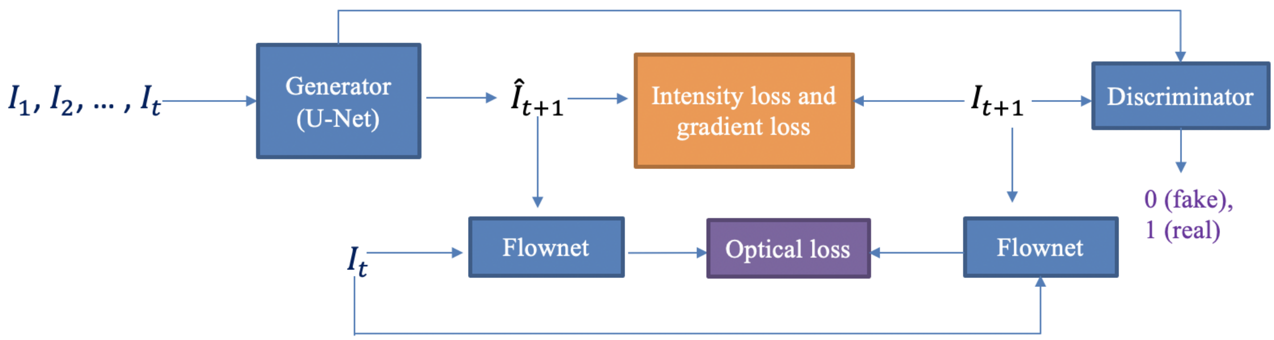

Figure 1. The pipeline of the video frame prediction we will use. We choose U-Net as generator to predict next frame. We adopt two kinds of constraints, appearance (intensity loss and gradient loss) and motion (optical flow). The Flow-net is the pretrained network for calculating optical flow. We use adversarial training to decide whether the prediction is real or fake

to detect deviations from learned rules. They typically perform linear approximations by PCA , clustering [4], non-linear approximation by various types of auto-encoders (VAE), Long Short Term memory networks (LSTMs) [21], and finally deep generative models (GANs). If the reconstruction error is high for a specific future instance, then it will be identified as an anomaly. The drawback of such method is that it does not train on anomalies. So there is no guarantee successful predictions for future anomalies. Also due to the high capacity of deep learning based methods, we could get low reconstruction errors for anomalies. There are also few supervised detection methods, where the normal profile is built using labeled data. It is typically applied for classification and regression related applications. [7].

Based on features used, all existing methods can be categorized into three categories: Hand-crafted features based, Deep learning based, and Frame Prediction Based.

2.1. Hand-crafted Features Based Method

Hand-craft features based method mainly include three steps. First, extracting features, features are either hand-crafted or learned from training dataset. Second, we train a model to find the distribution of regular scenarios. Third, we identify outliers or isolated clusters as anomalies. For feature extraction, early endeavors normally.

Early research normally utilizes low-level trajectory features, a sequence of coordinates to represent normal patterns in images. Yang et al. [14] propose a new feature descriptor called Mutli-scale Histogram of Optical Flow (MHOF). It describes motion direction information and preserves more accurate motion energy information. Cewu et al. proposed sparse combination learning for detection for high structure redundancy in surveillance videos. It directly learn sparse combinations to enhance testing speed. However, such methods do not perform well in complicated scenes with shadows and obstructions since trajectory features based on object tracking likely fail in these scenarios. Considering shortcoming of trajectory based features, histogram of oriented gradients (HOG) [2]and histogram of oriented flows (HOF) [3] are adopted. Grauman and Kim [8] modeled the distribution of local nodes' typical optical flow with a mixture of probabilistic principal component analyzers (MPPCA). Zhang et al. [23] proposed a semi-supervised adapted hidden Markov Model (HMM) framework based on spatial-temporal features. Adam et al. used multiple local monitors to describe the local histograms of optical flow by an exponential distribution. Apart from these statistical models, dictionary learning is another popular method to encode the normal patterns. The basic premise is that regular pattern itself can be described as a linear combination of basis of encoding dictionary. Hence, a pattern is identified as anomaly if the reconstruction error is high. However, in general, hand-crafted based methods can't perform well with noise information and in extreme environment such as the bad weather.

2.2. Deep Learning Based Anomaly Detection

Deep learning methods have already demonstrated successes in many computer vision domains as well as anomaly detection. Xu et al. [22] proposed Appearance and Motion DeepNet (AMDN) which utilizes deep neural networks to automatically learn feature representations. In another work, Mahmudul et al. [6] propose a 3D convolutional auto-encoder (Conv-AE) to model regular frames and learn both the local features and the classifiers as an end-to-end learning framework. Convolutional Neural Networks (CNN) is very capable of learning spatial features, while Recurrent Neural Networks (RNN) and Long Short Term Memory are always used to model sequential data. Hence, researchers leveraged both advantages of CNN and RNN. Jefferson and Andreas [17] proposed end-to-end trainable composite Convolutional Long Short Term Memory (Conv-

LSTM) networks to model both normal patterns and motion patterns same time, predicting the evolution of a video sequence from a small number of input frames. Luo et al. [15] proposed temporally coherent sparse coding (TSC) method, which can be mapped to a stacked RNN to facilitate the parameter optimization and accelerate the anomaly prediction.

Nonetheless, due to the high learning capacity and generality of deep learning based algorithms, there is not guarantee that we will successfully identify anomalies since we might not observe big difference in reconstruction errors between normal and abnormal events. Hence, this results in less efficient discrimination.

2.3. Future Frame Prediction Based Anomaly Detection

In recent years, prediction learning attracts increasingly attention from researchers for its potential application in unsupervised feature learning for patterns representation. In [16], a multi-scale network with adversarial training is proposed to generate more coherent and natural future video frames. In [20], shi et al. proposed the convolutional LSTM (ConvLSTM) and used it to build an end-to-end trainable model for the precipitation nowcasting problem. In [13], a predictive neural network is proposed 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. All these above works try to predict future frames directly. More recently, Liu et al. [12] proposed a new baseline for video anomaly detection, the future frame prediction based framework for anomaly detection. Other than just enforcing predicted frames to be close to the ground truth spatially, they also enforce optical flow between predicted frames to be close to their optical flow ground truth.

Figure 2. Traffic cameras usually have fixed angles so that they will capture similar scenes, which makes this task possible to solve. We applied a video anomaly detection method based on future frame prediction to detect traffic accidents.

3. Method

It will be nature to detect anomaly since anomaly can be adverse but normal events are stable. We choose a Closed World Assumption for our method, which means all events that have large distance between our predicted result will be considered as an anomaly. The whole process can be divided into two tasks: **Future Frame Prediction Task** and **Anomaly Detection Task**. Therefore, we need to train a generator to predict future frame based on historical observation for the Future Prediction Task and a detection metric for Anomaly Detection Task.

3.1. Future Frame Prediction Task

Wei and Weixin devised a pipeline to train a generator by Generative adversarial networks(GAN)[11]. Their work mostly focused on pedestrians on street or in front of a store. For traffic accident, we believe that the patterns between pedestrians and cars can be similar. Therefore, we used similar pipeline for our Future Frame Prediction Task to train a generator.

3.1.1 Backbone for Frame Prediction

In order to predict future frame, the input of our model will be $I_1, I_2, ..., I_t$, which I_i denotes the i^{th} frame in t consecutive frames, and the output will be the future frame $\hat{I}_{t+1}$. U-Net[18] provides with a good architecture to meet our requirements. Therefore, we choose U-Net as our backbone for frame prediction.

3.1.2 Adversarial Training

Generative adversarial networks is a powerful tool in image and video generation. In order to get a higher quality prediction, we also use GAN to train our generator better, as shown in Figure. 1. GAN contains a generator G, which is the U-Net mentioned in section 3.1.1, and a discriminator D. The discriminator is a classifier to classify the input into real frame or a generated frame. G learns to generate more valid frames that are harder for D. Meanwhile, D learns to better discriminate the generated frames. In this "Adversarial training" process, both D and G will get better trained. In practice, we will train D and G in order:

1. **Training** D**.** The goal of training D is to train a classifier to classify original frames into class 1 and classify generated frames into class 0. When training D, the weights of G will be fixed. We choose Mean Square Error(MSE) loss function as the target function. Therefore, the adversarial loss for D is:

$$L^D_{adv}(\hat{I}, I) = \sum_{i,j} \frac{1}{2}(D(\hat{I})_{i,j} - 0)^2 + \sum_{i,j} \frac{1}{2}(D(I)_{i,j} - 1)^2 \tag{1}$$

where i,j denotes the spatial indexes.

2. **Training** G. The goal of training G is to generate frames such that these frames can't be discriminated by D and be classified to class 1. Similarly, we will also fix the weights of D when training G. The loss function for G is:

$$L_{adv}^{G}(\hat{I}) = \sum_{i,j} \frac{1}{2}(D(\hat{I})_{i,j} - 1)^2 \tag{2}$$

3.1.3 Intensity and Gradient Losses

Besides adversarial training, we still need some pixel-level features to make the prediction close to the ground truth. We use the l_2 distance between our prediction and the ground true in intensity space and in gradient space.

For intensity space, we just define the intensity loss function as:

$$L_{int}^{G}(\hat{I}, I) = \left\|\hat{I} - I\right\|_2^2 \tag{3}$$

For gradient space, we first define the gradient for two directions in a frame as:

$$\begin{aligned} G_x(I) &= \sum_{i,j} |I_{i,j} - I_{i-1,j}| \\ G_y(I) &= \sum_{i,j} |I_{i,j} - I_{i,j-1}| \end{aligned} \tag{4}$$

Then, the loss function for gradient loss will be:

$$\begin{aligned} L_{gd}^{G}(\hat{I}, I) =& |G_x(I) - G_x(\hat{I})| \\ &+ |G_y(I) - G_y(\hat{I})| \end{aligned} \tag{5}$$

3.1.4 Optical Flow Loss

A significant difference between image and video is that video provide more temporal information than image. For example, if a subtle change occurs in a predicted frame, the intensity loss and gradient loss can be small but it can have different optical flow, which can be considered a metric to evaluate motion between frames. In order to take good advantage the temporal information, we also introduce optical flow loss to get a better prediction.

Flownet[5] is a CNN based approach to estimate optical flow. We use pretrained Flownet for our motion estimation, which means all the parameters in Flownet will be fiexed. Therefore, the motion loss will be:

$$L_m^G = |f(\hat{I}_{t+1}, I_t) - f(I_{t+1}, I_t)| \tag{6}$$

where f means the Flownet.

3.1.5 Global Loss Functions

The role for different loss function can be different. For example, the adversarial loss can be less important than the intensity loss since the similarity of frames can depend more on intensity. Therefore, we add a corresponding hyperparameter for each loss function such that we can try different ratio of these features, so the final global loss function will be:

$$\begin{aligned} L^G =& \lambda_{int} L_{int}^G(\hat{I}_{t+1}, I_{t+1}) \\ &+ \lambda_{gd} L_{gd}^G(\hat{I}_{t+1}, I_{t+1}) \\ &+ \lambda_m L_m^G \\ &+ \lambda_{adv} L_{adv}^G(\hat{I}_{t+1}) \end{aligned} \tag{7}$$

For discriminator, there is only one loss function L_{adv}^D, so the global loss function for D will be:

$$L^D = L_{adv}^D(\hat{I}_{t+1}, I_{t+1}) \tag{8}$$

3.2. Anomaly Detection Task

After finish the future frame prediction task, we will consider whether our generator is well trained in this task. Therefore, we need to compare the input frames and the predicted frames. We will consider our predicted frames will happen if there is no anomaly. Higher distance between input frames and predicted frames will have higher possibility to be abnormal. In this step, we will use Peak Signal to Noise Ratio (PSNR) as our metric to evaluation "how normal" this frame is.

3.2.1 Peak Signal to Noise Ratio

A simple way to compare the predicted frame and the real event is to compute the MSE of each pixel in each channel. However Mathieu[16] shows PSNR is a better way to evaluate the similarity between frames. PSNR is defined as:

$$PSNR(\hat{I}, I) = 10log_{10} \frac{[max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i - \hat{I}_i)^2} \tag{9}$$

3.2.2 Score for Detection

For a test video, we want to give a uniform standard to judge different videos. Therefore, for each video, we will compute a score for each frame to give the final evaluation for this whole video. The score is a normalized value in range [0, 1] and the score for the t^{th} frame in a video is defined as:

$$S(t) = \frac{PSNR(\hat{I}_t, I_t) - min_t PSNR(\hat{I}_t, I_t)}{max_t PSNR(\hat{I}_t, I_t) - min_t PSNR(\hat{I}_t, I_t)} \tag{10}$$

According to the settings of Wen[11], we can judge whether a frame is normal or abnormal based on its score. A threshold, for example 0.5, can be set to be the lowest score to be considered as normal.

Figure 3. Some samples in different scenes in CADP

4. Experiments

We apply our method on a traffic camera dataset to test the performance. This section will introduce our dataset, show the training details, and the result for prediction and detection.

4.1. Dataset

We use Car Accident Detection and Prediction (CADP) as our target dataset. Some samples are shown in Fig. 3.

CADP provides with different scenes with at least one anomaly. Each scene includes a video recorded by a fixed angle traffic camera, which means the normal frames will be similar so that it will be easier for our generator to predict.

Since our model only learns normal frames, we split one video into 4 parts: 1 abnormal part and 3 normal parts. 1 normal part and the abnormal part will be split into test set while the other two parts will be split into train set.

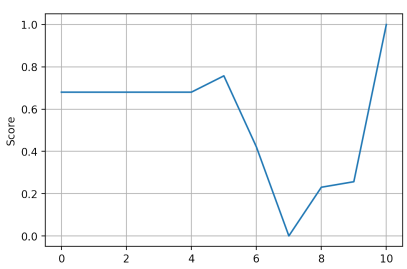

Figure 4. Score of each frame in a normal video clip.

4.2. Training Details

Our generator and discriminator are trained on Nvidia GeForce 3080 GPU for 3000 epochs. We use Adam[9] as our optimizer, the hyperparameters $\lambda_{int}, \lambda_{gd}, \lambda_m, \lambda_{adv}$ are set as 1.0,1.0,2.0, and 0.05 , and the learning rate for generator and discriminator are set as 0.001 and 0.00001 according to Wen's recommendation[11]. Each frame will be resized to 256×256, and we use sequence length 4 for our input.

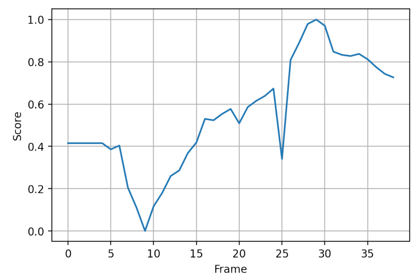

Figure 5. Score of each frame in an abnormal video clip.

4.3. Result for Prediction and Detection

We test on one scene on CADP due to time limitation. This subsection will show the result for prediction and detection.

4.3.1 Future Frame Prediction Task

Our result is visualized in Tensor Board[1]. Figure. 6 shows the prediction result in Future Frame Prediction Task. The left part is the ground truth and the right part is the predicted frame. For the static part such as road and stores, they are well predicted, but for moving objects, there are some subtle differences since the model is not well trained. But in global, this result is acceptable.

4.3.2 Anomaly Detection Task

For anomaly detection task, we compute the score $S(t)$ for each frame and set the abnormal threshold as 0.5, which means it will have a high probability for a frame whose score is lower than 0.5. Due to the normalization, even normal frame can have a low score in a video clip, but its pattern will be different from the abnormal frames. Here is an example:

1. In Figure. 4, a score for normal video clip is shown. We can see most frames are in a high level, only a few

Figure 6. Prediction result from generator. Left part is the ground truth in test set and the right part is the predicted frame.

frames are lower than the threshold.

2. In Figure. 5, a score for abnormal video clip is shown. We can see the first half video is in a low level. We check the original video, and find it is because of a motorcycle is driven over speed, which will cause a big difference.

4.4. Shortcomings of the Method

Although the model gives an acceptable result, there are still several shortcomings for this method:

1. **Model need to be well trained.** Our detection is highly based on the accuracy of the generator. Therefore, we need to train the model with adequate information such as different vehicle types and different colors.

2. **Model for different scene need to be trained individually.** The generator is not universal. Therefore, the model trained in scene A can't be used in scene B.

5. Conclusion

We propose the future frame prediction network for anomaly detection for traffic accidents. The model is trained by normal frames and tested by both normal and abnormal frames. Particularly, we use a U-Net as the base network. We impose a temporal loss to ensure the optical flow of predicted frame to be consistent with the ground truth and intensity and gradient losses to make sure that pixel-level features of the prediction close to the ground truth. Adversarial Training method ensures that Generator generated predicted frames are classified to class 1 (original frame) by discriminator. In this way, we generate normal events in terms of both motion and appearance and frames with higher quality prediction. Frames with large difference between prediction and ground truth would be identified as anomalies. Extensive experiments on CADP dataset show our method outperform existing methods by a large margin, which proves the efficiency of our selected algorithm.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[3] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer, 2006.

[4] Keval Doshi and Yasin Yilmaz. Fast unsupervised anomaly detection in traffic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 624–625, 2020.

[5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[6] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[7] Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, and Masao Sakauchi. Traffic monitoring and accident detection at intersections. *IEEE transactions on Intelligent transportation systems*, 1(2):108–118, 2000.

[8] Jaechul Kim and Kristen Grauman. 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. IEEE, 2009.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2):36, 2018.

[11] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline. 2017.

[12] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[13] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[14] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[15] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[16] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[17] Jefferson Ryan Medel and Andreas Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[19] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9. IEEE, 2018.

[20] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28:802–810, 2015.

[21] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *International Conference on Learning Representations*, 2018.

[22] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[23] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan. Semi-supervised adapted hmms for unusual event detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 611–618. IEEE, 2005.