A Little Improvement on Future Frame Prediction for Anomaly Detection Baseline

张国栋, 吕凤玲, 林逸泰

张国栋:36220201154084@stu.xmu.edu.cn, 吕凤玲:31520211154012@stu.xmu.edu.cn,
林逸泰:23020211153896@stu.xmu.edu.cn

Abstract

In recent years, with the widespread deployment of surveillance cameras in public places, abnormal detection in videos has attracted more and more research attention. The task of video anomaly detection is could identify unexpected behavior from a video stream is to identify events that rarely or never occur in a particular context. There are few detection methods for crowd gathering abnormal behavior in public places, and among all existing methods, reconstruction or future frame prediction is the chief method for video anomaly detection. The strong generalization ability of the reconstruction model, abnormal behaviors are often ignored. Due to the defects of the future frame prediction method itself, the robustness of the detection model is low, and there will be a false alarm for normal behavior. Therefore, we firstly replicated the classic Future Frame Prediction Base model in the CVPR video anomaly detection task in 2018, and then analyzed the existing problems through experiments, considering the combination of reconstruction and Future Frame Prediction to balance the shortcomings of existing models. A reconstruction module would be added to the baseline model, and expect experiments on benchmark datsets could show that the improved method is better than the previous model.

Introduction

With the development of the social economy, there are more and more cities with large-scale and super-large populations(Pang et al. 2021), which makes the issue of urban public safety more and more important. For example, the stampede incidents on the Bund and Hajj in Mecca were all due to abnormal crowd behaviors that were not detected and dealt with in time, resulting in serious group safety incidents (Xu et al. 2017). The population density in public areas such as subways, stations, and hospitals in large cities is relatively high. It is of great significance to ensure its normal order. In the above-mentioned specific places and sensitive areas, abnormal behaviors often occur between crowd activities, which may cause serious threats to the personal safety of the crowds. Therefore, timely detection and treatment are required. When solving anomaly detection tasks, it is generally believed that anomalous events should be unexpected events that occur less frequently. But this has caused another difficulty. In real scenarios, both normal and abnormal behaviors are diverse and undefinable. Therefore, the detection model trained in the above manner lacks good robustness in practical applications. To solve the above problems, the predecessors have done a lot of work. Among them, most of the methods (Cong, Yuan, and Liu 2011; Hasan et al. 2016; Luo, Liu, and Gao 2017) apply the idea of reconstruction, that is, minimize the reconstruction error of the training data (normal behavior) during training, and the reconstruction error can be compared with the test during testing. The deep neural network (self-encoder) has strong learning generalization ability; the reconstruction method does not consider the semantic information of the video context, that is, the generated reconstructed feature map is completely based on its single frame. In recent years, with the development of GAN (Goodfellow et al. 2020), the performance and effect of video frame prediction (Mathieu, Couprie, and LeCun 2015) have also been greatly improved. Therefore, an anomaly detection method based on future frame prediction (Liu et al. 2018) has also been proposed.

The method based on frame prediction can overcome the limitation based on the reconstruction method. The information between the predicted frames must be input when the model generates the prediction. To ensure the quality of the predicted frame image, the model often quotes the method based on frame prediction. Enter the optical flow information representing the motion feature as the motion constraint between the generated frame and the predicted frame (Liu et al. 2018). But using optical flow, the model will be extremely sensitive to changes in lighting in the application scene, and it is prone to false reports. In addition, the prediction frame generated by the method based on frame prediction is highly dependent on the input previous frame information, so the model detection result may be very sensitive to any changes in the input previous frame, resulting in low robustness in practical applications.

In summary, we will first reproduce the classic Future Frame Prediction Base model (Liu et al. 2018) in the task of video anomaly detection, analyze its specific problems through experiments, and finally improve the model algorithm. Specifically, by adding a reconstruction module, trying to combine the respective advantages of the above two methods and balance their existing shortcomings, without reducing the accuracy of model anomaly detection, improve

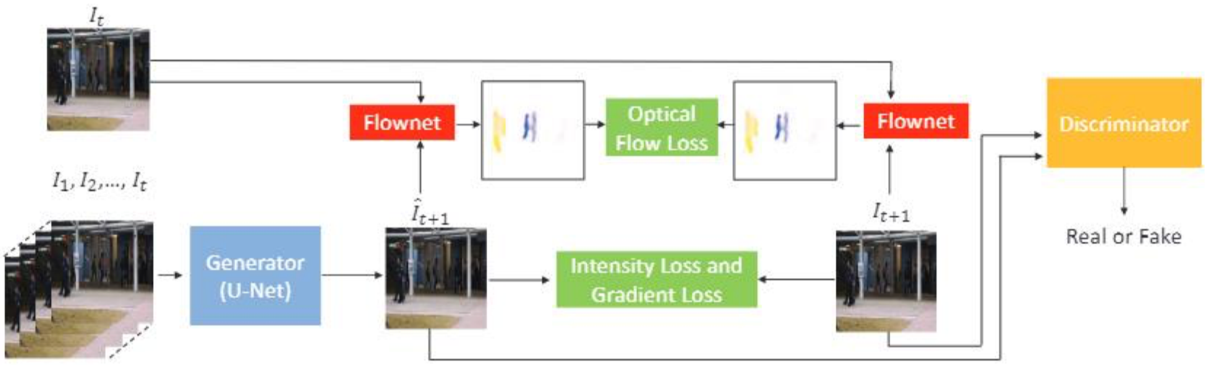


Figure 1: Future Frame Prediction Model. The U-Net like generator generates the predicted frame, and the intensity loss and spatial loss is calculated based on the intensity and gradient information. The FlowNet is hired to calculated the temporal loss.

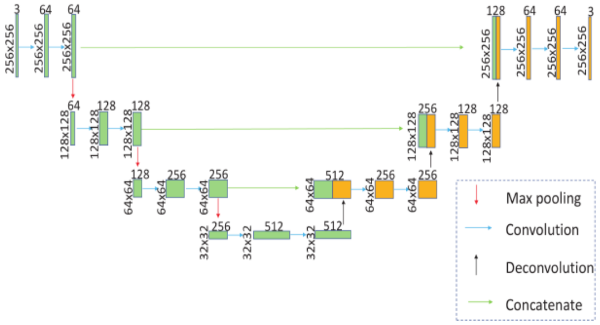


Figure 2: The U-Net architecture. Our model utilized U-Net to overcomes the gradient-vanishing problems and information imbalance of corresponding layers of the encoder and the decoder.

the model's robustness in practical applications.

Related Works

2.1 Anomaly detection based on manual features Anomaly detection based on manual features includes manual feature extraction, learning model establishment and anomaly cluster detection distinguish. The commonly used manual feature extraction methods are: histogram of oriented Gradients(Dalal and Triggs 2005), histogram of optical flow(Dalal, Triggs, and Schmid 2006), 3D gradient, local binary pattern (LBP). Common learning models and discrimination methods include: clustering normal samples by using common clustering algorithms, such as k-means, k-medoids, fuzzy c-means and Gauss hybrid model, and then treating the points far away from the clustering center in the test set as abnormalities. In addition to these statistical models, sparse coding or dictionary learning(Zhao, Fei-Fei, and Xing 2011) are also commonly used to encode normal behavior patterns. The underlying assumption of these methods is that normal behavior patterns can be well linearly encoded into a representational dictionary, while abnormal patterns can not be well represented and detected.

2.2 Anomaly detection based on deep learning ConvLSTM(Xingjian et al. 2015) is a good example, which has been proposed to solve the problem of precipitation prediction. Most anomaly detection methods are based on self encoder. The self encoder can reconstruct the original data through unsupervised learning, so as to learn the efficient representation of the input data. The development of generation countermeasure network not only promotes the development of high-quality generation tasks, but also promotes the development of video anomaly detection method based on frame prediction. The high variance within positive samples, and the inherent data-imbalance problem(Chalapathy and Chawla 2019) of the video anomalies. (Nayak, Pati, and Das 2021). AEs can often start reconstructing anomalies as well which depletes their anomaly detection performance. To mitigate this, (Astrid, Zaheer, and Lee 2021) propose a temporal pseudo anomaly synthesizer that generates fake-anomalies using only normal data. (Dong, Zhang, and Nie 2020) predict future frames for normal events via a generator and attempt to force the predicted frames to be similar to their ground truths. MIST(Feng, Hong, and Zheng 2021) is composed of a multiple instance pseudo label generator, which adapts a sparse continuous sampling strategy to produce more reliable clip-level pseudo labels, and a self-guided attention boosted feature encoder that aims to automatically focus on anomalous regions in frames while extracting task-specific representations.

Future Frame Prediction Model

Future Frame Prediction for Anomaly Detection - A New Baseline (Liu et al. 2018) firstly utilized the difference of predicted frame and ground truth frame for abnormal detection in videos. Except for the intensity and gradient information of the generated predicted frame added as spatial constraints, the temporal constraint is added by focusing the optical flow of the predicted frame and the corresponding ground truth frame to be the same. This leads to higher quality predicted frame for normal events, as shown in Fig.1.

The Generator

The general architecture of a frame-generating network consists of two modules: an encoder,which extracts the feature frame-by-frame by reducing the spatial resolution, and a decoder,which recovers the spatial resolution with corresponding layers. However, such solutions suffer from great gradient-vanishing problems and information imbalance of corresponding layers of the encoder and the decoder. Our model utilized U-Net (Ronneberger, Fischer, and Brox 2015) as a generator to solve this problem. It was initially proposed as a convolutional neural network for biomedical image segmentation, and it adopts a skip connection structure between upper and lower sampling layers of the same resolution. This method can not only effectively solve the problem of model training gradient disappearance and the unbalance of each layer of coding or decoding information, but also is very suitable for video anomaly detection. Because the surveillance video has a constant foreground and only sensitive moving crowd activities change, u-NET can be used to restore the foreground of the video image so that model training is more focused on crowd activities. The detailed structure of the generator U-net is shown in Fig. 2.

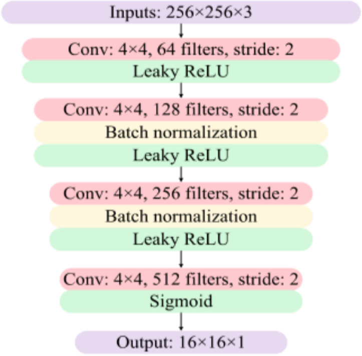

Inputs: 256×256×3
Conv: 4×4, 64 filters, stride: 2
Leaky ReLU
Conv: 4×4, 128 filters, stride: 2
Batch normalization
Leaky ReLU
Conv: 4×4, 256 filters, stride: 2
Batch normalization
Leaky ReLU
Conv: 4×4, 512 filters, stride: 2
Sigmoid
Output: 16×16×1

Figure 3: The Patch GAN discriminator. Each element of the output matrix is mapped to the corresponding patch in a frame, and the value of the patch is used to judge whether the patch contains exceptions.

The Discriminator

For better predictions of the approximate position of abnormal events in the video frame image, PatchGAN () 's discriminator was used as the discriminator of the model. The main difference from the discriminator of traditional GAN network lies in that the latter maps the input image to a single scalar output within the range of [0,1], representing the probability of whether the image is true or false, while PatchGAN provides a matrix as the output, with each element representing whether the corresponding patch is true or false. Fig. 3 shows the details of PatchGAN. Each element of the output matrix is mapped to the corresponding patch in a frame, and the value of the patch is used to judge whether the patch contains exceptions.

Constrains

In order to make the quality of the generated prediction frame close to the real frame, correlation constraints need to be added. Following the work of () in generating prediction frames, pixel intensity and gradient changes are also used as constraints. The intensity constraint, which minimizes the L2 distance between the generated prediction frame and the real frame, guarantees the similarity of all pixels in the RGB space of the two frames. However, adding this constraint is not enough, and the generated prediction frame is prone to ambiguity and other problems. The gradient change constraint ensures that the gradient between each pixel in the generated prediction frame and its horizontal and vertical pixels is consistent with the situation in the real frame, which can sharpen the generated prediction frame image well. The intensity and gradient constraints are formulated as follows, where i and j represent the spatial index of video frames:

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

$$\begin{aligned} L_{gd}(\hat{I}, I) = \sum_{i,j} & \| |\hat{I}_{i,j} - \hat{I}_{i-1,j}| - |I_{i,j} - I_{i-1,j}| \|_2^2 + \\ & \| |\hat{I}_{i,j} - \hat{I}_{i,j-1}| - |I_{i,j} - I_{i,j-1}| \|_2^2 \end{aligned} \tag{2}$$

Not only the quality problems of spatial intensity and gradient can be predicted, but also the correctness of motion prediction of prediction frame expression can be guaranteed as much as possible. Therefore, the approximate time (motion) constraint is defined by the optical flow estimation between the predicted frame and the real frame, that is, the motion state of the predicted frame is consistent with that of the real frame. Flownet(), which realizes optical flow estimation based on CNN, is used to calculate optical flow. The constraint formula is as follows, where F represents the calculation of optical flow function

$$L_{op}(\hat{I}, I) = \|f(\hat{I}_{t+1}, I_t) - f(I_{t+1}, I_t)\|_1 \tag{3}$$

During training, The purpose of discriminator D is to classify real frames as 1 and predicted frames as 0. When discriminator D is trained, the weight of generator G is adjusted and the loss function of mean square error (MSE) is specified as follows, where I and j represent the spatial index of prediction frame patch:

$$\begin{aligned} L_{adv}^{D}(\hat{I}, I) = & \sum_{i,j} \frac{1}{2} L_{MSE}(D(I)_{i,j}, 1) \\ & + \sum_{i,j} \frac{1}{2} L_{MSE}(D(\hat{I})_{i,j}, 1) \end{aligned} \tag{4}$$

When training generator G, the weight of discriminator D is fixed, and the mean square error function is also used. The specific formula is as follows, where I and j represent the spatial index of prediction frame patch:

$$L_{adv}^{G}(\hat{I}) = \sum_{i,j} \frac{1}{2} L_{MSE}(D(\hat{I})_{i,j}, 1) \tag{5}$$

The objective function of the model is determined by introducing the previously mentioned constraints. The loss function of generator G is shown as follows, where $\lambda_{int}, \lambda_{gd}, \lambda_{op}, \lambda_{adv}$ are set as 1.0, 1.0, 2.0, 0.05.

$$\begin{aligned} L_G = & \lambda_{int} L_{int}(\hat{I}_{t+1}, I_{t+1}) \\ & + \lambda_{int} L_{gd}(\hat{I}_{t+1}, I_{t+1}) \\ & + \lambda_{int} L_{op} \\ & + \lambda_{int} L_{adv}^{G}(\hat{I}_{t+1}) \end{aligned} \tag{6}$$

And the loss of the discriminator is:

$$L_D = L_{adv}^D(\hat{I}_{t+1}, I_{t+1}) \tag{7}$$

Abnormal detection

On the standard of anomaly detection, the method based on frame prediction idea is grasped, that is, the concept that normal events can be predicted while abnormal events cannot be predicted is followed. To put it simply, normal events can be generated well, while abnormal events can be generated at lower quality. Therefore, anomaly detection can be carried out by taking advantage of the quality difference between pre-measured frames and real frames. In terms of the method to measure the quality of the predicted image, the peak signal-to-noise ratio (PSNR) is adopted following the work of (), and the formula is as follows:

$$PSNR(I, \hat{I}) = 10 \times \log_{10} \frac{max(\hat{I})^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i - \hat{I}_i)^2} \tag{8}$$

In order to better select the threshold value for evaluating anomalies, PSNR is normalized, and then the abnormal detection can be carried out by setting the appropriate threshold value according to the experiment. The normalization function is shown as follows:

$$S(t) = \frac{PSNR(I_t, \hat{I}_t - min_t(PSNR(I_t, \hat{I}_t)))}{max_t(PSNR(I_t, \hat{I}_t)) - min_t(PSNR(I_t, \hat{I}_t))} \tag{9}$$

Improved algorithm

According to the previous work, we reproduce this method,and we use it for video exceptions with public datasets of the detection task (described in detail in the experimental part),such as Avenue, USCDped2 ,and acquire good results. However, we have been tried in RWF2000 and other anomaly detection data that are biased towards real scenes,but the results are extremely unsatisfactory, which makes us have to think about the reason why this method can't work.Through testing and analysis, we find some defects.Firstly, the anomaly detection based on frame prediction emphasizes that abnormal events cannot be detected, but it ignores that many normal events are unpredictable, so the false alert rate is high.Secondly,the model adds a Flownet module and uses it to calculate the optical flow, trying to use the optical flow estimation to introduce the time (motion) characteristics.

Although most anomaly detection methods introduce optical flow information to improve the accuracy of the model, the results show that the use of optical flow information can indeed introduce more features for detection to a certain extent.However, in addition to character motion, the change of light also has a great impact on optical flow estimation. Especially at night or in dark scenes, the change of light is easy to cause drastic changes in optical flow, resulting in wrong detection.Moreover, when reproducing the method, we try to remove the Flownet module (removing the optical flow constraint), but it is found that the experimental results are not greatly reduced, only one or two points fluctuate, but the efficiency of the overall model is improved. Therefore, it is necessary to consider the optical flow information for the model structure of the algorithm.Thirdly, the model is easily affected by noise, but the surveillance video in the real scene often contains large noise, so the effect of the model is poor.To sum up, let's give a simple example to better explain the existing problems.For example, in Figure 5, in the night monitoring scene, the sudden lighting of the lamp is a normal event, but the PSNR difference between the two possible future frames (the lamp is on or not on) is too large, resulting in misjudgment.

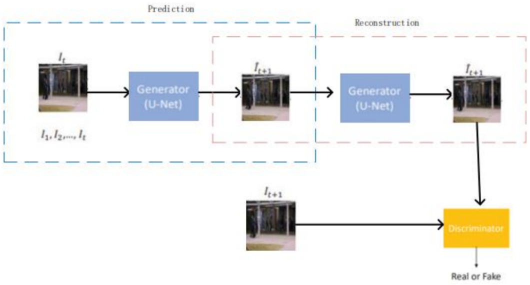


Figure 4: Improved future prediction model framework

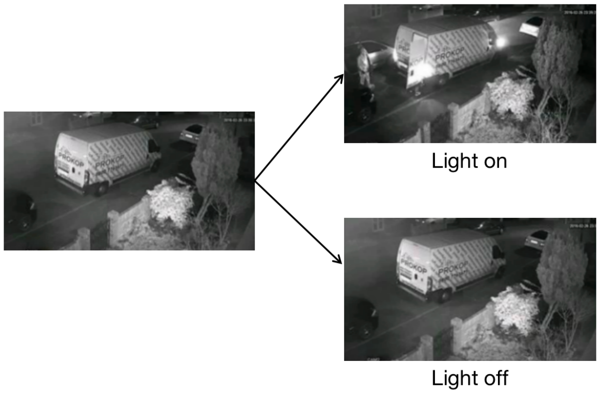


Figure 5: Two prediction results of sudden lighting of lights at night.

In order to reduce the impact of video noise on model performance and reduce the unpredictability of normal events as much as possible,we learn from the anomaly detection method based on reconstruction(Cong, Yuan, and Liu 2011; Hasan et al. 2016; Luo, Liu, and Gao 2017).However, in order to keep the overall model framework unchanged as much as possible and reduce the adjustment of the model structure as much as possible, the solution is to add a reconstruction module to the frame prediction model, and the reconstruction module uses the structure adjusted U-net.What's more, in order to make the model applicable to a wider range of scenarios, reduce the complexity of the model and increase the real-time prediction, we remove the Flownet module.The overall improved model is shown in Figure 4.

5.1 Dataset

UCSD dataset.UCSD is the pedestrian dataset, which consists of Ped1 and Ped2.Ped1 contains 34 training videos and 36 test videos with a frame resolution of 238 × 158 pixels. Ped2 contains 16 training videos and 12 test videos with a frame resolution of 360 × 240 pixels. Ped1 and ped2 have different perspectives, including bicycles, vehicles, skateboards and wheelchairs passing through the pedestrian area. CUHK Avenue dataset.It contains 16 training videos and 21 test videos, including 15328 frames in the training set and 15324 frames in the test set. Anomalies include throwing objects, wandering and running. For each test frame, frame level exception annotation is used. The resolution of each video frame is 360 × 640 pixels.

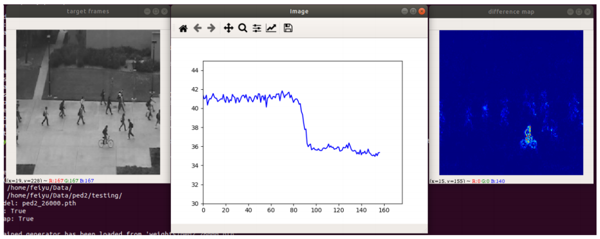
Experiment and analysis

5.2Evaluating indicator

Because the above data sets have the following characteristics: the training set is all normal behavior, and the test set has many normal behavior and very few abnormal behavior. Therefore, the problems faced make the positive and negative samples of the data extremely unbalanced, the evaluation index does not use the algorithm accuracy, but uses the area under curve (AUC).

5.3Experimental results

We set the hyperparameters of improved model by referring to the hyperparameters of Future Frame Prediction Basemodel, use the same optimizer and adjust the learning rate. For comparative experiments, we set the same batch size and the number of training rounds.Experiments are conducted and compared on avenue and UCSD Ped2 . The experimental results are shown in the figure 7,figure **??**, and Table1.(The results are all recurrence, Future Frame Prediction Basemodel's results deviate from the original paper, but not much)



```
(x=19, y=228) ~ R:167 G:167 B:167
 /home/feiyu/Data/
 /home/feiyu/Data/ped2/testing/
del: ped2_26000.pth
: True
ap: True

ained generator has been loaded from 'weights/ped2_26000.pth
```

Figure 6: Model demo display diagram (including real-time real frame, real-time prediction frame, PSNR change, difference depth diagram between real frame and prediction frame)

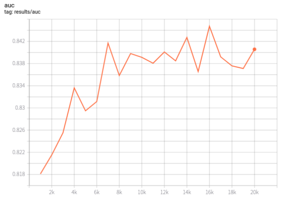


Figure 7: Results of the improved model on the avenue dataset (20000 rounds).

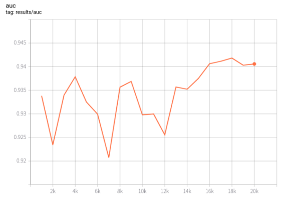


Figure 8: Results of improved model on UCSD Ped2 dataset (20000 rounds)

Table 1: Comparison of results of Future Frame Prediction Basemodel and the improved model on two data sets.

	avenue	***USCD ped2***
future predict	83.7%	95.3%
future predict + reconstruction	84.6%	96.2%

Conclusion

This paper mainly analyzes and summarizes the video anomaly detection tasks and their difficulties, introduces the common methods in this task and their advantages and disadvantages, and focuses on video anomaly detection based on frame prediction method.During the experiment of reproducing and trying other data sets, we find the existing problems, so we try to introduce the reconstruction method to solve them, and finally achieve good results.Unfortunately, all the improvements we have made are to monitor the application of video in real scenes, namely increasing the robustness of the algorithm in practical applications as much as possible, but there is no experiment on the data set in line with the relevant scenes.The reasons are that there are some surveillance video datasets in real scenes, but the quality of the datasets is poor and the pre-processing is difficult; The performance of the monitoring video dataset of the model in the real scene fluctuates greatly, which may be because the monitoring video dataset is rich in scenes and abnormal behavior elements, so it is difficult to make comparison.Finally, we would like to thank the teacher for his teaching in the course of deep learning. We have learned a lot about deep learning, and we also thank teaching assistants for their help.

References

Astrid, M.; Zaheer, M. Z.; and Lee, S.-I. 2021. Synthetic Temporal Anomaly Guided End-to-End Video Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 207–214.

Chalapathy, R.; and Chawla, S. 2019. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Cong, Y.; Yuan, J.; and Liu, J. 2011. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, 3449–3456. IEEE.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 886–893. Ieee.

Dalal, N.; Triggs, B.; and Schmid, C. 2006. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, 428–441. Springer.

Dong, F.; Zhang, Y.; and Nie, X. 2020. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8: 88170–88176.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14009–14018.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 733–742.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545.

Luo, W.; Liu, W.; and Gao, S. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, 341–349.

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

Nayak, R.; Pati, U. C.; and Das, S. K. 2021. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106: 104078.

Pang, G.; Shen, C.; Cao, L.; and Hengel, A. V. D. 2021. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2): 1–38.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810.

Xu, M.; Li, C.; Lv, P.; Lin, N.; Hou, R.; and Zhou, B. 2017. An efficient method of crowd aggregation computation in public areas. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10): 2814–2825.

Zhao, B.; Fei-Fei, L.; and Xing, E. P. 2011. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, 3313–3320. IEEE.