Anomaly Detection of Predicted Frames Based on U-Net Feature Vector Reconstruction

Yong Qiang¹, Shumin Fei*¹, Yiping Jiao¹ and Liuwen Li²

¹ Key Laboratory of Measurement and Control of CSE (Ministry of Education), School of Automation, Southeast University, Nanjing, 210096, China
² School of Computer Engineering, Jinling Institute of Technology, Nanjing, 211169, China;

*Corresponding author: smfei@seu.edu.cn

Abstract: Anomaly detection in surveillance video scenes is one of the current research hotspots. Due to the small sample collection of anomalous events, the lack of negative sample labeling data training in anomaly detection research adds a lot of difficulties. Therefore, we adopt the method of unsupervised training and improve the method of anomaly detection based on the reconstruction of the potential features of the predicted frame and ground truth based on u-net. We reduce the reconstruction error between the potential features of u-net in the predicted frame and the potential features of the real frame. Then through other constraints, the reconstruction error of the entire predicted frame is minimized according to the generative adversarial training. Due to the use of normal behavior sample training, when the abnormal behavior is detected, the reconstruction error value exceeds the set threshold to judge whether abnormal behavior occurs in the surveillance video. Experiments prove that our improved method is effective and accurate.

1. Introduction
Anomaly detection based on surveillance video scenes is an important research field. In many studies[1], deep learning algorithms have been widely used, and this method has the characteristics of high accuracy. Most methods of monitoring video anomaly detection based on deep learning still have supervised learning training[2], that is, it takes a lot of time and manpower to label data and long-term training. The normal behavior of the video in the surveillance scene still occupies the vast majority, and there are few samples of abnormal behavior. This results in a mismatch of the normal distribution of positive and negative samples, and the consistency of supervised learning and training cannot be achieved. Therefore, how to design a video anomaly detection model for unsupervised learning is important to research work.

In this framework, the following two models are trained simultaneously: a generator model G that captures the data distribution and a discriminator model D that estimates the samples in the probability of the training data[3]. The main idea is that G is to generate real images as far as possible to deceive D, and D is to judge the images generated by G as false. G and D make the model achieve the best purpose of generating images through the game. Now, this method has been widely used and developed.

In[4], the author proposed an anomaly detection method based on video prediction framework, and also cited some constraints to keep the prediction frame consistent with ground truth. In[5], the author
proposes a convolutional neural network structure with an encoder and decoder with jump connections and cites the adversarial training method. The reconstruction error of the latent vector of the image is minimized to achieve the normal distribution required for model learning.

Since normal events are predictable and abnormal events are unpredictable, it is feasible to construct a framework for detecting abnormal video behaviors based on predicted frames under such an idea. In this article, inspired by the above literature, we combined with the existing methods for improvement. The video frame is predicted through the U-net structure in order to make the predicted frame closer to the ground truth. The latent feature vector obtained by the encoder of the real frame and the latent feature vector of the predicted frame in U-net is approximated and minimized. At the same time, context constraints and SPynet (This method calculates optical flow through the spatial pyramid and deep learning network.)[6] optical flow constraints are also required for the predicted and real frames. Under such an anomaly detection model based on U-net latent feature vector reconstruction, we conducted experiments on some public data sets to verify the effectiveness of our method.

This paper contributes (1) to improve the method of anomaly detection based on the U-net prediction frame and ground truth latent feature reconstruction. (2) Using efficient SPynet optical flow constraints. The detection accuracy of the entire prediction frame anomaly detection frame is improved. (3) The method has good generalization ability and can be applied to abnormal behavior detection in some low-density pedestrian monitoring scenarios.

2. Related work

2.1. Deep learning methods
Among the anomaly detection methods of surveillance video, anomaly detection methods based on deep learning are widely used. For example, the literature [7] adopts the method of 3D convolutions to detect anomalies in video. The spatial and temporal features are extracted through 3D convolutions, and a loss constraint with decreasing weights is used to generate predicted frames.

2.2. GAN
The proposal of GAN theoretical methods has aroused the attention and research of many scholars. Among them, in video anomaly detection, literature [4] is based on the theoretical method of GAN and constitutes a framework for generating adversarial networks based on video prediction frames. Reference [8] builds an anomaly detection model of encoder-decoder-encoder sub-network in GAN.

2.3. Latent feature vector
The BIGANs proposed in [9] prove that GAN can learn inverse mapping and project the data back to the latent space vector. It shows that the obtained latent feature learning is useful.

2.4. Prediction frame
Literatures [4] and [7] are based on predicting future frames of video frames and take the reconstruction error of the predicted frames as the detection object. Set a threshold to determine whether abnormal behavior occurs. Because the experiment of the entire model is based on the training of normal event samples. Therefore, if abnormal behavior occurs in the detection target, its reconstruction error will be greater than the set threshold.

3. Anomaly detection model based on U-net latent feature vector reconstruction
Abnormal behaviors in monitoring scenes are mostly behavioral events that are very different from normal behaviors. And the number of samples of abnormal behavior events is very small, which is far inferior to the number of samples of normal behavior events. In this paper, abnormal behaviors are judged according to constraints such as the reconstruction of potential feature vectors of predicted frames and a series of learning to generative adversarial training.
As shown in Figure 1, the entire anomaly detection framework includes a prediction frame module, a latent feature vector reconstruction module, a constraint module, and a generation confrontation training module. Using the video frame sequence frame $F=(F_1, F_2, L, F_t)$ as an input frame, a frame prediction frame $\hat{F}_{t+1}$ is generated through U-net reconstruction. $\hat{F}_{t+1}=G_e(F)$, and $o$ is the latent feature vectors of the input frame. In order to ensure that more global multi-scale information is retained during the reconstruction of the predicted frame, encoder $\hat{G}_e$ is used to compress the original video $F_{t+1}$ to obtain the potential feature vector $\hat{o}$. $\hat{o}=\hat{G}_e(F_{t+1})$. Here we need to make the dimensions of $\hat{o}$ and $o$ the same, and minimize the parameter distance between them to minimize the reconstruction error.

![Figure 1. Anomaly detection frame structure of predicted frame based on latent feature vector](image)

**Figure 1.** Anomaly detection frame structure of predicted frame based on latent feature vector

U-net’s network structure reference[4,10], $\hat{G}_e$ adopts the structure of DCGAN[8,11]. The structure of $\hat{G}_e$ is shown in Figure 2. In the predicted frame module, the reconstruction of the latent feature vector is adopted to restrict the generation of the predicted frame. We have also adopted two other constraint methods: Intensity Loss and optical flow loss. This is to ensure that the space prediction and motion estimation can be accurately predicted in video prediction. Finally, through discriminator D, the real future frame $F_{t+1}$ is classified into class 1, and B is classified into class 0. Among them, class 1 and class 0 respectively represent true labels and false labels.

**3.1. Feature Loss**

In order to make $\hat{F}_{t+1}$ and $F_{t+1}$ consistent, we minimize the potential space vectors compressed by the video frames of both. $\hat{o}$ and $o$ can be expressed as $\hat{o}=\hat{G}_e(F_{t+1})$ and $o=G_e(F)$.

$$L_{FE} = \|G_e(F) - \hat{G}_e(F_{t+1})\|_2$$  

(1)
3.2. Intensity Loss
In order to further constrain the accuracy and consistency of the reconstruction of the predicted frame, $\hat{F}_{t+1}$ and $F_{t+1}$ are constrained by $L_1$, and $L_1$ is used instead of $L_2$ because $L_1$ can reduce the ambiguity [12].

$$L_{IL} = \| F_{t+1} - \hat{F}_{t+1} \|_1$$ (2)

3.3. Optical Flow Loss
Under the previous constraints, the motion information of the moving target in the video cannot be accurately estimated, so optical flow loss is added [4]. In this paper, the method of SPynet [6] is used to estimate the optical flow, and the optical flow estimation formula is defined as $f_{\theta}(y)$.

$$L_{OF} = \| f_{\theta}(F_{t+1}, F_t) - f_{\theta}(\hat{F}_{t+1}, F_t) \|_1$$ (3)

3.4. Adversarial Loss
In generative adversarial training learning, the $\hat{F}_{t+1}$ generated by the prediction module $G$ is close to $F_{t+1}$ to the greatest extent, while the Discriminator classifies the generated samples and the real samples. The purpose of training is to reconstruct the predicted frame to the maximum extent close to the real sample. Make $G$ and $D$ complete the minimum and maximum game.

The goal of training $D$ is to classify $A$ as labels 1 and $B$ as labels 0 [4,13].

$$L_{advD} = \sum_{i,j} \left( \frac{1}{2} | \hat{Z} - Z | (D(F)_{i,j}, 1) + \frac{1}{2} | \hat{Z} - Z | (D(F)_{i,j}, 0) \right)$$ (4)

And $\hat{Z} \in [0,1]$, $Z = 1$ or $Z = 0$.

The goal of training $G$ is to enable $D$ to classify $\hat{F}_{t+1}$ as label 1.

$$L_{advG} = \sum_{i,j} \frac{1}{2} | \hat{Z} - Z | (D(\hat{F})_{i,j}, 1)$$ (5)

3.5. Combining Losses
Combine the above loss functions as follows:

$$L_G = \lambda_{FE} L_{FE}(\hat{F}_{t+1}, F_{t+1}) + \lambda_{IL} L_{IL}(\hat{F}_{t+1}, F_{t+1}) + \lambda_{OF} L_{OF}(\hat{F}_{t+1}, F_{t+1}) + \lambda_{adv} L_{advD}(\hat{F}_{t+1})$$ (6)

$$L_D = L_{advD}(\hat{F}_{t+1}, F_{t+1})$$ (7)

In the above formula, $\lambda_{FE}$, $\lambda_{IL}$, $\lambda_{OF}$ and $\lambda_{adv}$ are weight parameters, which are the effects of each loss function on the objective function.

3.6. Evaluation Calculation
After training the anomaly detection model, the detected frame $a$ is compared with the predicted frame $\hat{a}$ to perform anomaly detection. Using the method of [14], the following score formula is obtained.

$$A(a, \hat{a}) = \lambda L_G(a, \hat{a}) + (1 - \lambda) L_D(a, \hat{a})$$ (8)

Where $L_G(a, \hat{a})$ is the reconstruction error value of Formula (2), which is the Intensity loss between the detected frame and the predicted frame, and $L_D(a, \hat{a})$ is the reconstruction error of the potential feature vector of the detected frame and the predicted frame, which is the reconstruction error value of Formula (1). $\lambda$ is the weight parameter of the calculation formula. Normalize the $A(a, \hat{a})$ of
the detected video frame to $[0,1]$, and then obtain the score of the detected frame according to the following formula.

$$E(t) = 1 - \frac{A(a_i, \hat{a}_i) - \min A(a_i, \hat{a}_i)}{\max A(a_i, \hat{a}_i) - \min A(a_i, \hat{a}_i)}$$

(9)

4. Experiments

4.1. Datasets

UCSD Ped1. The dataset contains 34 training samples and 36 test samples. The abnormal behavior in the video includes cart, wheelchair, biker among pedestrians.

UCSD Ped2. The dataset contains 16 training samples and 12 test samples. Its abnormal behavior is the same as Ped1.

CUHK Avenue. The dataset contains 16 training samples and 21 test samples. There are three types of abnormal behavior: abnormal object, strange action, and wrong direction.

ShanghaiTech Campus dataset. The dataset consists of 13 different surveillance scenarios with over 130 abnormal behaviors.

4.2. Preparation for training experiments

Before training the anomaly detection model, we set the size of all video frames to $256 \times 256$, while $T=1$, crop 5 consecutive frames. The learning rate is set to $e^{-9}$ and the weight parameters are set to $\lambda_{ff}=1, \lambda_{li}=0.6, \lambda_{of}=1$ and $\lambda_{adv}=0.05$.

4.3. Anomaly detection experiment

Before the test samples are tested, the training samples are first trained for learning. Because the training and learning samples are all normal behaviors, the reconstruction error of the predicted frame and the training sample is almost the same. When the trained model is tested on the test sample, if the calculated regularity score is lower than the set threshold, it can be judged that the test sample has an abnormal event.

As shown in Figure 3, the trained anomaly detection model is used for detection experiments. When pedestrians walk correctly in the monitoring area, their predicted frames can be predicted normally. When abnormal behavior occurs, the predicted image is blurred and there is a large reconstruction error.

In four public datasets, the designed prediction frame anomaly detection model is tested. The experimental results obtained are shown in Figure 4. The yellow area represents the abnormal behavior of the test video. The result proves that the predicted frame anomaly detection model can effectively detect anomalies. It can be seen from Table 1 that our method has a better anomaly detection effect than most other methods.
We refer to the previous work[15][16] and change the different thresholds to calculate the AUC value of ROC to evaluate the performance of the model.

![Scoring graphs for anomaly detection in four public datasets.](image)

**Figure 4.** Scoring graphs for anomaly detection in four public datasets.

**Table 1.** Comparison of AUC values in different public data sets with other anomaly detection methods.

| Method       | UCSD Ped1 | UCSD Ped2 | CUHK Avenue | ShanghaiTech |
|--------------|-----------|-----------|-------------|--------------|
| Unmasking[17]| 68.4      | 82.2      | 80.6        | None         |
| sRNN[18]     | None      | 92.2      | 81.7        | 68.0         |
| Conv-AE[15]  | 81        | 90        | 70.2        | None         |
| FFPAD[4]     | 83.1      | 95.4      | 85.1        | 72.8         |
| Our method   | 86.9      | 96.1      | 86.2        | 74.1         |

5. Conclusions

We improved an anomaly detection model for predicted frames based on U-net latent feature vector reconstruction. Experiments on public datasets prove that our improved method is superior to most existing anomaly detection methods. And the flexibility and accuracy are very high. In future research work, we will collect new datasets for detection and adjust more complex surveillance scenarios. And study the new GAN method and integrate different loss functions in the anomaly detection framework to study whether there is a good improvement effect.
References

[1] Sabokrou M, Fayyaz M, Fathy M, Moayed Z and Klette R 2018 Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes *Comput. Vis. Image Underst.* **172** 88–97

[2] Di Mattia F, Galeone P, De Simoni M and Ghelfi E 2019 A Survey on GANs for Anomaly Detection. http://arxiv.org/abs/1906.11632

[3] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in neural information processing systems* pp 2672–80

[4] Liu W, Luo W, Lian D and Gao S 2018 Future Frame Prediction for Anomaly Detection - A New Baseline *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 6536–45

[5] Akcay S, Atapour-Abarghouei A and Breckon T P 2019 Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection *Proceedings of the International Joint Conference on Neural Networks* vol 2019-July (IEEE) pp 1–8

[6] Ranjan A and Black M J 2017 Optical flow estimation using a spatial pyramid network *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* 2017-Janaua 2720–9

[7] Zhao Y, Deng B, Shen C, Liu Y, Lu H and Hua X S 2017 Spatio-temporal AutoEncoder for video anomaly detection *MM 2017 - Proc. 2017 ACM Multimed. Conf.* 1933–41

[8] Akcay S, Atapour-Abarghouei A and Breckon T P 2018 Ganomaly: Semi-supervised anomaly detection via adversarial training *Asian Conference on Computer Vision* (Springer) pp 622–37

[9] Donahue J, Darrell T and Krähenbühl P 2019 Adversarial feature learning *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*

[10] Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **9351** 234–41

[11] Radford A, Metz L and Chintala S 2016 Unsupervised representation learning with deep convolutional generative adversarial networks *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*

[12] Isola P, Zhu J Y, Zhou T and Efros A A 2017 Image-to-image translation with conditional adversarial networks *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* 2017-Janaua 5967–76

[13] Mathieu M, Couprie C and LeCun Y 2016 Deep multi-scale video prediction beyond mean square error *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*

[14] Zenati H, Foo C S, Lecouat B, Manek G and Chandrasekhar V R 2018 Efficient GAN-Based Anomaly Detection. https://arxiv.org/pdf/1802.06222

[15] Hasan M, Choi J, Neumann J, Roy-Chowdhury A K and Davis L S 2016 Learning temporal regularity in video sequences *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem* 733–42

[16] Schlegl T, Seeböck P, Waldstein S M, Schmidt-Erfurth U and Langs G 2017 Unsupervised anomaly detection with generative adversarial networks to guide marker discovery *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **10265** LNCS 146–7

[17] Ionescu R T, Smeureanu S, Alexe B and Popescu M 2017 Unmasking the Abnormal Events in Video *Proc. IEEE Int. Conf. Comput. Vis. 2017-Octob* 2914–22

[18] Luo W, Liu W and Gao S 2017 A revisit of sparse coding based anomaly detection in stacked rnn framework *Proceedings of the IEEE International Conference on Computer Vision* pp 341–9