A Promotion Method for Generation Error Based Video Anomaly Detection

Zhiguo Wang, Zhongliang Yang, Yu-Jin Zhang

Abstract—Using the generation error (GE) of a generative neural network (GNN) to detect video anomalies exhibits excellent performance. However, there are two problems when using the trained GNN models to detect anomalies. First, utilizing the frame-level GE to detect anomalies reduces the anomaly saliences, because anomalies usually occur in local areas. Second, when multiple discriminants (a discriminant is an anomaly score sequence) are available, using the weighted sum method to aggregate multiple discriminants does not always perform effectively, and the weights are hard to tune. To address these problems, we propose an approach consists of two modules. Firstly, we replace the frame-level GE with the maximum of the block-level GEs in the frame to detect anomalies. Secondly, assuming that the higher the anomaly threshold, the more reliable the anomaly detected, we propose a reliable-anomaly (R-anomaly) based strategy to aggregate multiple discriminants. We use the R-anomalies in the auxiliary discriminants to enhance their anomaly scores in the main discriminant. Experiments are carried out on UCSD and CUHK Avenue datasets. The results demonstrate the effectiveness of the proposed method and achieve state-of-the-art performance.

Index Terms—anomaly detection, anomaly saliency, multiple discriminants aggregation, surveillance video

I. INTRODUCTION

SURVEILLANCE videos are important for social security. However, it is time-consuming and labor-intensive to watch the surveillance videos for a long time, as most of the content of the videos are normal. The task of video anomaly detection is to detect anomalies automatically in surveillance videos. It is challenging, because the anomalies rarely happen and the types of anomalies are uncountable.

Currently, the video anomaly detection algorithms can be summarized into two classes: the traditional machine learning based methods [1]–[11] and the deep-learning based methods [12]–[22].

The traditional machine learning based methods usually use the hand-crafted-features [1]–[7] or the deep-features [8]–[10] to construct feature space, and then utilize the traditional machine learning models to detect the outliers. The commonly utilized models include: hidden Markov model (HMM) [1][2], sparse coding [3]–[6], topic model [7] and one-class support vector machines (OC-SVM) [8]–[11].

The deep-learning based methods leverage the neural networks to learn the manifold distribution of the normal samples, and then classify the samples deviated from this distribution as anomalies. Among the deep-learning based methods, many works [12]–[19] utilize a GNN to reconstruct/predict a frame, and then utilize the reconstruction/prediction error to detect anomalies. Reconstruction error and prediction error are both generation errors. In this letter, we call such methods GE-based methods.

The GE-based methods show excellent performances in video anomaly detection. Hasan et al. [12] utilized the reconstruction error of the autoencoder (AE) to detect anomalies. To capture more temporal information, Chong et al. [15] utilized the long short-term memory autoencoder (LSTM-AE) to reconstruct the input frames, and utilized the reconstruction error to detect anomalies. Luo et al. [16] utilized the LSTM-AE to reconstruct the current frame and the previous frame at each time step, then utilized the weighted sum of these two reconstruction errors to detect anomalies. Lee et al. [17] utilized a bidirectional LSTM-AE to synthesize an inter-frame in a sequence, besides that, they used a 3D convolutional discriminator to determine whether the generated sequence is real. They combined the GE and the loss of discriminator to detect anomalies. Liu et al. [18] utilized the U-net to predict the future frame, and used multiple losses as the constraint of the output, including: pixel loss, gradient loss, flow loss and discriminator loss. They utilized the pixel loss to detect anomalies.

These methods focus on the optimization of the GNN models. After the model been trained, they employed the similar strategy to detect anomalies: They utilized the frame-level GEs to calculate anomaly scores and then to detect anomalies [12]–[18]. When multiple discriminants were available, they usually utilized one of them [13][14][18] or utilized their weighted sum to detect anomalies [16][17].

However, there are two problems in this strategy. (1) The GE existed in every position in the frame, the anomalies usually only occur in local areas. The frame-level GE calculated the GE of the whole scene that reduced the anomaly saliences. (2) Using the weighted sum method to combine multiple discriminants does not always improve the performance, as it combines both the advantages and the disadvantages of different discriminants. Moreover, the weights are hard to learn and cannot be adjusted according to the prior knowledge.

To address these problems, this letter proposes an approach consists of two modules. (1) We replace the frame-level GE with the maximum of the block-level GEs in the frame to detect anomalies. (2) We propose a R-anomaly based aggregation method to combine multiple discriminants: Firstly, with the help of the training data, we set a strict anomaly detection
threshold (SADT) for every auxiliary discriminant and utilize it to detect R-anomalies. Then we count the frequencies that the test frames be classified as R-anomalies (R-freqs) and use them to enhance the anomaly scores in the main discriminant. Finally, we use the enhanced main discriminant to detect anomalies. The framework of the proposed approach is shown in Fig. 1.

where $A$ is the number of pixels on $l$, $e_l$ is the GE value of the $i$-th normal pixel on the normal/abnormal frame, $e'_l$ is the GE value of the $j$-th anomaly pixel on the abnormal frame, and $n$ is the number of abnormal pixels on the abnormal frame.

We can assume that the GE of the normal area in the abnormal frame is equal to the GE of the corresponding normal area in the normal frame. Then the anomaly saliency can be written as:

$$\text{Saliency} = \frac{\sum_{i=1}^{n} e'_i - \sum_{i=1}^{n} e_l}{\sum_{i=1}^{n} e_l + \sum_{i=1}^{n} e'_i}$$  \hspace{1cm} (5)$$

From equation (5), we can find that the anomaly saliency is negatively correlated to $\sum_{i=1}^{n} e_l$, given the specific abnormal frame. That is, the higher the contribution of the GE of the normal regions to the anomaly score, the lower the saliency of the abnormal frames.

Therefore, in order to improve the anomaly saliencies of the abnormal frames, we reduce the contribution of the GE of the normal regions to the anomaly score. We utilize the block-level GE to detect anomalies: We first put a size-fixed sliding window on the frame, and term the GE of a window area as block-level GE. Then, we calculate the block-level GE at each window position. Finally, we select the maximum of the block-level GEs on the frame to detect anomalies.

$$L_{B_k} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} B_{k,i,j}$$ \hspace{1cm} (6)$$

$$L_B = \max\{L_{B,1}, ..., L_{B,K}\},$$ \hspace{1cm} (7)$$

where $B_k$ is the $k$-th block on the frame, $K$ is the total number of the blocks on a frame, $h$ and $w$ are the height and width of the sliding window, respectively. $L_{B_k}$ is the block-level GE of $B_k$, $L_B$ is the maximum of the block-level GEs on the frame. Note that the block-level operation can be implemented by a convolution layer for mean filtering and a max-pooling layer. Therefore, this operation can be accelerated on the GPU. Because the GEs have noises along the time axis, we use the median filter to filter out the noises after this operation.

B. Anomaly Score

To calculate the anomaly score, many works [12][18] used the max-min normalization strategy to normalize the GEs in each video. The reason for this process is that different videos have different normal-GE-levels, and the difference between different normal-GE-levels may be larger than that caused by anomalies.

However, this strategy has a problem: it produces high anomaly scores in every video, even if there is no anomaly in the video. This is not in line with the application needs of our real life.

Considering the shortcoming of the normalize strategy and that the block-level operation improves the anomaly saliency, we use the $L_B$ without normalization as the anomaly score to detect anomaly.

$$s(t) = L_B(t),$$ \hspace{1cm} (8)$$

C. Aggregate Multiple Discriminants

Let $\{S_1, ..., S_D\}$ be $D$ discriminants of a video, where $S_t = [s_t(1), ..., s_t(t), ..., s_t(T)]$, $s_t(i)$ is the $i$-th anomaly score in the $t$-th discriminant, $T$ is the number of the frames...
of the video. Firstly, we choose one of the discriminants as main discriminant $S_{\text{main}}^t$, and treat the remaining discriminants as auxiliary discriminants $S_{1}^t, ..., S_{t-1}^t$, $S_0$. Then, we aggregate multiple discriminants as follows:

1) We calculate SADT for each auxiliary discriminant $S_i^t$, and use it to detect R-anomalies:

\[
\hat{s}_i^t(\alpha) = \beta \cdot \hat{s}_i^t(\alpha \cdot N),
\]

where $\beta$ is a parameter to adjust the strictness of the SADT.

2) With the help of the calculated SADT, we detect R-anomalies in each auxiliary discriminant.

\[
f_i(t) = \begin{cases} 
1, & s_i(t) \geq \text{SADT}_i \\
0, & s_i(t) < \text{SADT}_i 
\end{cases}
\]

where $f_i(t)$ is the R-anomaly label for the $t$-th test frame under the discriminant $S_i^t$.

3) We use the detected R-anomalies to enhance the corresponding anomaly scores in the main discriminant.

\[
f(t) = \sum_{i=1}^{D} f_i(t),
\]

where $f(t)$ is the R-freq of the $t$-th frame in all auxiliary discriminants.

3.2) We smooth the R-freqs along the time axis. Assuming that an anomaly event lasts at least $\varepsilon$ frames [12], the frame between two anomaly frames should also be an anomaly, if the time distance between the two anomaly frames is smaller than $\varepsilon$. With this assumption, we smooth the R-freqs with the following formula:

\[
f_i'(t) = \min(f_i, f_b), \quad \text{if } f_i \leq \min(f_a, f_b), a < t < b, b - a \leq \varepsilon.
\]

3.3) We use the smoothed R-freqs to enhance the corresponding anomaly scores in the main discriminant $S_{\text{main}}^t$.

\[
S_{\text{main}}^t(t) = S_{\text{main}}^t(t) \cdot (\gamma \cdot f_i'(t) + 1),
\]

where $\gamma$ is a parameter to adjust the impact of $f_i'(t)$ to $S_{\text{main}}^t(t)$.

In this method, we have 4 parameters: $\alpha, \beta, \gamma, \varepsilon$. Each parameter has its specific physical meaning. Therefore, we can adjust them according to the application requirements and the prior acknowledge.

III. EXPERIMENTS

Work [18] is a typical GE-based method. We refer to the method proposed in [18] to evaluate the proposed approach.

A. Dataset and Evaluation Criteria

The experiments are carried out on two datasets: CUHK Avenue dataset [3] and UCSD Pedestrian dataset [23]. The Avenue dataset contains 16 training videos and 21 testing videos. The abnormal events include running, throwing schoolbag, throwing papers, etc. The UCSD dataset has two sub-datasets: Ped1, Ped2. The two sub-datasets capture different scenarios but have the similar definition of abnormal events, include cycling, skateboarding, crossing lawns, cars, etc. These two sub-datasets are usually used separately.

The most commonly used evaluation metric is the Receiver Operation Characteristic (ROC) and the Area Under Curve (AUC). Following the work [18], we detect the frame-level anomalies and use the frame-level AUC for performance evaluation.

B. Effectiveness of Block-level Operation

We analyze the effectiveness of the block-level operation from three perspectives: the anomaly saliency, the AUC of anomaly detection and the impact of the block-size to the AUC.

From the GNN model mentioned in [18], we can generate three GE maps: the GE map of the pixel value $GE_{\text{pixel}}$, the GE map of the optical flow $GE_{\text{flow}}$, and the GE map of the gradient value $GE_{\text{grad}}$.

Fig. 2 visualizes the intensity map and the heat map of the $GE_{\text{pixel}}$ of a frame in Ped2. As the figure shows, the GE exists in almost all areas where the foreground targets are located. The GE of the normal area is strong enough to weaken the anomaly saliency in the anomaly scores generated by the frame-level GE.

Fig. 2. Visualization of the $GE_{\text{pixel}}$ map on Ped2. (a) The raw frame. (b) The intensity map of $GE_{\text{pixel}}$. (c) The heat map of $GE_{\text{pixel}}$.

Fig. 3 shows the anomaly scores (the higher the anomaly saliency, the better the anomaly score) calculated with the frame-level $GE_{\text{pixel}}$ and the block-level $GE_{\text{pixel}}$. As shown in the figure, the anomaly saliency is improved significantly by using the block-level GEs, and the saliency is significant enough to resist the interference of the difference between different normal-GE-levels on abnormal detection.

Fig. 3. Anomaly score curves in frame-level and block-level $GE_{\text{pixel}}$ of Ped2, both without normalization. The frames in the red areas are anomalies. (a) frame-level. (b) block-level, $h = w = 30$.

Table I and Table II list the impact of the block-level operation to the anomaly saliences and that of the AUCs, respectively. As shown in the two tables, the block-level operation can improve AUCs and anomaly saliences significantly on multiple GEs multiple datasets. It proves the correctness of our analysis and the effectiveness of the block-level operation.

Fig. 4 shows the influence of the block-size to the AUC. With the increase of the block-size, AUCs first increase and then
decrease. There is a large valid candidate interval to set the block-size to achieve better performance than that of the frame-level GE.

| Table I | The impact of block-level process to the anomaly saliences on multiple datasets and multiple GE |
|---------|------------------------------------------------------------------------------------------------|
| Saliency | Ped1 | Ped2 | Avenue |
| GE_pixel | Frame-level | 1.1047 | 0.6250 | 2.6341 |
| Block-level | 2.4979 | 2.0874 | 3.4552 |
| GE_flow | Frame-level | 1.0903 | 0.7023 | 2.4865 |
| Block-level | 1.6225 | 1.3034 | 2.7777 |
| GE_gdt | Frame-level | 0.1128 | 0.1131 | 0.2965 |
| Block-level | 0.4442 | 0.6189 | 0.7185 |

| Table II | The impact of block-level process to AUCs on multiple datasets and multiple GE |
|---------|------------------------------------------------------------------------------------------------|
| GE | AUC | Ped1 | Ped2 | Avenue |
| GE_pixel | Frame-level | 0.7946 | 0.8688 | 0.8766 |
| Block-level | 0.8291 | 0.9530 | 0.8982 |
| GE_flow | Frame-level | 0.7832 | 0.8401 | 0.8058 |
| Block-level | 0.8625 | 0.9712 | 0.8607 |
| GE_gdt | Frame-level | 0.7097 | 0.7618 | 0.7518 |
| Block-level | 0.7784 | 0.9642 | 0.8420 |

Fig. 4 The impact of the block-size to the AUC

C. Effectiveness of Multiple Discriminants Aggregation

Using GE_pixel, GE_flow, GE_gdt, we can get three GE-based discriminants: S_pixel, S_flow, S_gdt. From the discriminator of the model [18] we can get another discriminant S_adv. Without the loss of generality, we choose S_pixel as the main discriminant to evaluate the effectiveness of the aggregation method.

Fig. 5 shows several results from the aggregation process. In this process, we set: α_flow = α_gdt = α_adv = 0.01, β_flow = β_gdt = 1.2, β_adv = 0.99, ε = 50 [12], γ = 1. As shown in the Fig. 5(a-c), most of the detected R-anomalies are true anomalies, that proves the credibility of the assumption for R-anomalies. As shown in the Fig. 5(g), the anomaly scores of the R-anomaly frames are enhanced, which increases the likelihood that these video frames will be detected as anomalies.

Table III shows the AUCs of multiple discriminants and that of the enhanced main discriminant S_main'. The parameters are same to that in Fig.5, except γ = 4. The enhanced main discriminant achieves much better performance than the raw main discriminant. It demonstrates that the aggregation method is effective.

D. Comparison with Existing Methods

The comparison of our method with other methods on UCSD Pedestrian and CUHK Avenue datasets is shown in Table IV.

![Fig. 5. Process of multiple-discriminants-aggregation. The frames in the red areas are anomalies. (a-c) The R-anomalies detected in S_flow, S_gdt, S_adv, respectively. (d) The R-freq. (e) The smoothed R-freq. (f) The raw anomaly scores in main discriminant S_pixel. (g) The aggregated anomaly scores S_main' (green area), compared with the raw anomaly scores S_pixel (blue area).](image)

![Fig. 4. Impact of the block-size to the AUC.](image)

Fig. 4. The impact of the block-size to the AUC.

### Table I

| Loss | Saliency | Ped1 | Ped2 | Avenue |
|------|----------|------|------|--------|
| GE_pixel | Frame-level | 1.1047 | 0.6250 | 2.6341 |
| Block-level | 2.4979 | 2.0874 | 3.4552 |
| GE_flow | Frame-level | 1.0903 | 0.7023 | 2.4865 |
| Block-level | 1.6225 | 1.3034 | 2.7777 |
| GE_gdt | Frame-level | 0.1128 | 0.1131 | 0.2965 |
| Block-level | 0.4442 | 0.6189 | 0.7185 |

### Table II

| GE | AUC | Ped1 | Ped2 | Avenue |
|----|-----|------|------|--------|
| GE_pixel | Frame-level | 0.7946 | 0.8688 | 0.8766 |
| Block-level | 0.8291 | 0.9530 | 0.8982 |
| GE_flow | Frame-level | 0.7832 | 0.8401 | 0.8058 |
| Block-level | 0.8625 | 0.9712 | 0.8607 |
| GE_gdt | Frame-level | 0.7097 | 0.7618 | 0.7518 |
| Block-level | 0.7784 | 0.9642 | 0.8420 |

### Table III

| AUC | Ped1 | Ped2 | Avenue |
|-----|------|------|--------|
| S_flow | 0.8625 | 0.9712 | 0.8607 |
| S_gdt | 0.7784 | 0.9662 | 0.8420 |
| S_adv | 0.5397 | 0.6668 | 0.8441 |
| S_pixel | 0.8291 | 0.9530 | 0.8982 |
| S_main' | 0.8524 | 0.9889 | 0.9173 |

### Table IV

| Method | Ped2 | Avenue |
|--------|------|--------|
| GMFC-VAE [22] | 92.2 | 83.4 |
| WTA-AE [8] | 96.6 | 82.1 |
| Conv-AE [12] | 85.0 | 80.0 |
| Conv-LSTM-AE [15] | 88.1 | 77.0 |
| Cross-channel [21] | 95.5 | N |
| STAN [17] | 96.5 | 87.2 |
| U-net [18] | 95.4 | 85.1 |
| U-net [18] (*) | 86.88 | 87.65 |
| U-net + block-level (*) | 95.30 | 89.82 |
| U-net + block-level + aggregate (*) | 98.89 | 91.73 |

Compared with the work [18], using the block-level operation and the aggregation operation, the performance is improved significantly. We achieve state-of-the-art performance on Ped2 and Avenue datasets.

IV. Conclusion

In this letter, we propose an approach to improve the performances of the GE-based algorithms. First, we use block-level GEIs to calculate anomaly scores to improve the anomaly saliency. Then, we propose a R-anomalies based method to aggregate multiple discriminants. The aggregate strategy can be extended to aggregate multiple discriminants generated from multiple models. We will do that in the future.
REFERENCES

[1] L. Kratz and K. Nishino, “Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models,” in 2009 IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1446–1453.

[2] S. Zhou, W. Shen, D. Zeng, and Z. Zhang, “Unusual event detection in crowded scenes by trajectory analysis,” in 2015 IEEE Int. Conf. Acoust. Speech Signal Process., 2015, pp. 1300–1304.

[3] C. Lu, J. Shi, and J. Jia, “Abnormal Event Detection at 150 FPS in MATLAB,” in Proc. IEEE Int. Conf. Comput. Vis., 2013, pp. 2720–2727.

[4] Y. Cong, J. Yuan, and J. Liu, “Abnormal event detection in crowded scenes using sparse representation,” Pattern Recognit., vol. 46, no. 7, pp. 1851–1864, Jul. 2013.

[5] H. Ren, W. Liu, S. I. Olsen, S. Escalera, and T. B. Moeslund, “Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection,” in Proc. Br. Mach. Vis. Conf. 2015, 2015, pp. 28.1-28.13.

[6] Y. Yuan, Y. Feng, and X. Lu, “Structured dictionary learning for abnormal event detection in crowded scenes,” Pattern Recognit., vol. 73, pp. 99–110, 2018.

[7] X. Hu, Y. Huang, X. Gao, L. Luo, and Q. Duan, “Squirrel-Cage Local Binary Pattern and Its Application in Video Anomaly Detection,” IEEE Trans. Inf. Forensics Secur., vol. 14, no. 4, pp. 1007–1022, Apr. 2019.

[8] H. Tran and D. Hogg, “Anomaly Detection using a Convolutional Winner-Take-All Autoencoder,” in Proc. Br. Mach. Vis. Conf. 2017, 2017.

[9] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, “Detecting Abnormal Events in Video Using Narrowed Normality Clusters,” in 2019 IEEE Winter Conf. Appl. Comput. Vis., 2019, pp. 1951–1960.

[10] D. Xu, Y. Yan, E. Ricci, and N. Sebe, “Detecting anomalous events in videos by learning deep representations of appearance and motion,” Comput. Vis. Image Underst., vol. 156, pp. 117–127, Mar. 2017.

[11] R. Hinami, T. Mei, and S. Satoh, “Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 3619–3627.

[12] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning Temporal Regularity in Video Sequences,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 733–742.

[13] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, “Spatio-Temporal AutoEncoder for Video Anomaly Detection,” in Proc. 25th ACM Int. Conf. Multimed., 2017, pp. 1933–1941.

[14] J. R. Medel and A. Savakis, “Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks,” arXiv:1612.00390, Dec. 2016.

[15] Y. S. Chong and Y. H. Tay, “Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder,” in arXiv:1701.01546, 2017, pp. 189–196.

[16] W. Luo, W. Liu, and S. Gao, “Remembering history with convolutional LSTM for anomaly detection,” in 2017 IEEE Int. Conf. Multimed. Expo, 2017, pp. 439–444.

[17] S. Lee, H. G. Kim, and Y. M. Ro, “STAN: Spatio-Temporal Adversarial Networks for Abnormal Event Detection,” in 2018 IEEE Int. Conf. Acoust. Speech Signal Process., 2018, pp. 1323–1327.

[18] W. Liu, W. Luo, D. Lian, and S. Gao, “Future Frame Prediction for Anomaly Detection - A New Baseline,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6536–6545.

[19] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, “Abnormal event detection in videos using generative adversarial nets,” in 2017 IEEE Int. Conf. Image Process., 2017, vol. 10878, pp. 1577–1581.

[20] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, “Adversarially Learned One-Class Classifier for Novelty Detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3379–3388.

[21] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, “Training Adversarial Discriminators for Cross-Channel Abnormal Event Detection in Crowds,” in 2019 IEEE Winter Conf. Appl. Comput. Vis., 2019, pp. 1896–1904.

[22] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine, “Video Anomaly Detection and Localization via Gaussian Mixture Fully Convolutional Variational Autoencoder,” arXiv:1805.11223, May 2018.

[23] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2010, pp. 1975–1981.