Multi-stage Transformer for Video Anomaly Detection

Viet-Tuan Le[1], Khuong G. T. Diep[1], Tae-Seok Kim[1], Yong-Guk Kim[1]
[1]Department of Computer Engineering, Sejong University, Seoul, Korea

tuanlv@sju.ac.kr, khuongdiep@sju.ac.kr, 22110617@sju.ac.kr, ykim@sejong.ac.kr

Abstract

Video anomaly detection aims to detect abnormal events. Motivated by the power of transformers recently shown in vision tasks, we propose a novel transformer-based network for video anomaly detection. To capture long-range information in video, we employ a multi-scale transformer as an encoder. A convolutional decoder is utilized to predict the future frame from the extracted multi-scale feature maps. The proposed method is evaluated on three benchmark datasets: USCD Ped2, CUHK Avenue, and ShanghaiTech. The results show that the proposed method achieves better performance compared to recent methods.

1. Introduction

Anomaly detection in video is an important task in surveillance video analysis, which aims to identify any unexpected events. However, video anomaly detection is a challenging problem in computer vision fields. The challenges of this task can be summarized in three major points. First, normal events are easy to collect, but anomaly patterns are difficult to describe. Moreover, anomaly samples are few and are acquired at a high cost. Second, the difference between a normal and an abnormal event is very close. For example, a walker and a skateboarder have similar appearance and velocity on crowded sidewalk. Third, realistic surveillance video data are complex and were recorded in different scenes. In general, an anomaly detection framework is trained on video data containing only normal events with an unsupervised setting.

Recently, unsupervised anomaly detection approaches can be divided into two categories: frame reconstruction [15] and future frame prediction [16], [12]. The frame reconstruction approaches often contain an encoder to generate a compressed encoding from the video frames and a decoder that reconstructs the corresponding output frames from the encoding. The reconstruction error between the generated frames and the target frames is used to determine whether the generated outputs are anomalous or not because the abnormal events correspond with larger reconstruction errors. The future frame prediction approaches assume that normal events can be well predicted and vice versa. A prediction network is trained to predict the future frame with high quality for normal training data. The difference between a predicted frame and its ground truth frame determines whether it is a normal or abnormal event.

In this study, we propose a novel network for video anomaly detection, which is based on transformer. The network consists of a transformer-based encoder and a convolutional decoder. The encoder is used to extract multi-scale feature maps, while the decoder is employed to predict the future frame from the extracted multi-scale feature maps.

The rest of this paper is organized as follows: Some related studies about anomaly detection in video are discussed in Section 2. The proposed network is presented in Section 3. In Section 4, experiments of the proposed network are discussed. Finally, Section 5 concludes this work.

2. Related work

Anomaly detection in videos has attracted a large number of researchers. Many works **Error! Reference source not found.**, [8], [9] proposed a modified U-Net-based prediction network wherein a visual relation module was added to exploit the visual relationship that was computed between the items in the foreground. In order to exploit the motion information, many studies [10], [11], [13] adopted 3D convolution that fused the spatial and temporal features simultaneously. Hao et al. [11] used a 3D CNN-based encoder and 2D CNN-based decoder as a generator to generate the future frame while Yang et al. [10], [13] predicted the future frames by using a 3D U-Net. Moreover, [13] proposed a patch anomaly generation by applying spatial rotation transformation and temporal mixing transformation to force the model to learn the spatial and temporal information of normal events. Chen et al, [14] proposed a bidirectional prediction network that includes a forward prediction network and a backward prediction network to predict the same target frame.

Both prediction and reconstruction networks had their disadvantages, Liu et al. [9] proposed a stack of a prediction network and a reconstruction network to complement each other. In another similar framework, Qiang et al. [17]

combined the reconstruction approach and the future frame prediction approach to exploit the advantages of both methods.

3. Method

In this section, we present the proposed method for video anomaly detection.

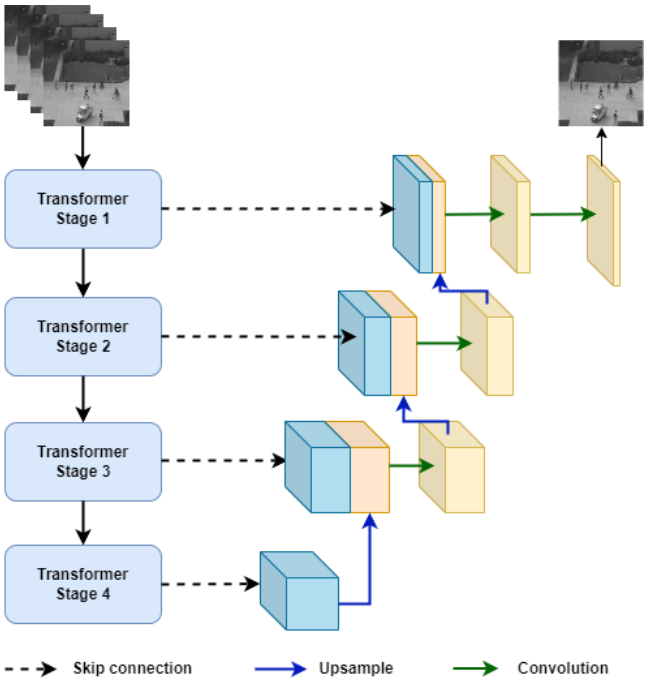


Figure 1. The overall architecture of the proposed method.

3.1. Overall Architecture

The overall architecture of the proposed method is illustrated in Figure 1, which includes a transformer-based encoder and convolutional decoder. The encoder generates hierarchical feature maps, while decoder predicts the future frame using the extracted features from the encoder.

3.2. Transformer-based Encoder

In order to generate multi-scale feature maps, we leverage Pyramid Vision Transformer [7] as the encoder. The encoder includes four stages, each stage generates a different scale. A stage consists of a patch embedding and several transformer layers. A transformer layer includes a spatial-reduction attention module and a feed-forward network. LayerNorm is placed before the attention module and the feed-forward network.

3.3. Convolutional Decoder

The extracted multi-scale features are fed into the decoder to generate the output frame. The decoder consists of several layers. First, the output from the last stage of the encoder is upsampled by using a deconvolutional layer. Then the resized features are combined with the output features of the encoder stage that have the same resolution via skip connection. The combined features are fed into several convolution layers, including a 3 x 3 convolution, batch normalization, and ReLU activation function. In the same way, using the output features from the previous layer as input, we can obtain the predicted future frame.

3.4. Objective function

Our network aims to predict the future frame I_t from a sequence of input frames $\{I_1, I_2, \dots, I_{t-1}\}$. To ensure the predicted frame close to its ground truth, we apply several loss functions.

To guarantee the similarity of the predicted frame and its ground truth, the l_2 loss is used:

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

In addition, a gradient loss is used to solve the drawback of the intensity loss.

$$L_{gra}(I,\hat{I}) = \sum_{i,j} \left\| |\hat{I}_{i,j} - \hat{I}_{i-1,j}| - |I_{i,j} - I_{i-1,j}| \right\|_1 + \left\| |\hat{I}_{i,j} - \hat{I}_{i,j-1}| - |I_{i,j} - I_{i,j-1}| \right\|_1$$

The final loss is computed as follows:

$$L_{pre}(I,\hat{I}) = \alpha L_{int}(I,\hat{I}) + \beta L_{gra}(I,\hat{I})$$

Where α and β are two coefficients that balance the weights of the loss functions.

3.5. Anomaly detection

PSNR is used to estimate the quality of the predicted frame.

$$PSNR(I,\hat{I}) = 10 log_{10} \frac{[max_I]^2}{\frac{1}{n}\sum_{i=1}^{N}(I_2 - \hat{I_2})^2}$$

The anomaly score of a frame is computed as follows:

$$S(t) = \frac{PSNR_t - \min(PSNR)}{\max(PSNR) - \min(PSNR)}$$

where min(PSNR) and max(PSNR) denote the minimum and the maximum PSNR values in the given video sequence, respectively

4. Experiments

4.1. Video anomaly detection datasets

The proposed network was evaluated on three benchmark anomaly detection datasets.

- UCSD Ped2 dataset contains 16 videos for training and 12 videos for testing, corresponding to 2550 frames for training and 2010 for testing, respectively. Ped2 contains 12 abnormal events which include cyclists, skateboarders, cars, people in wheelchairs, and people walking across a walkway or the grass.
- CUHK Avenue dataset consists of 16 videos for training and 21 videos for testing. It contains total of 30,652 frames which are split into 15,328 training frames and 15,423 testing frames. The resolution of each frame is 360x640 pixels. The dataset contains 47 abnormal events such as throwing objects, loitering, and running across the gate.
- ShanghaiTech Campus dataset was recorded in 13 different scenes with different light conditions and camera angles. It contains 437 videos and is split into 330 videos for training and 107 videos for testing. The training set contains 274,515 frames, which include normal events while the testing set contains 42,883 frames with 130 abnormal events.

Each video has a resolution of 480x856 pixels.

4.2. Comparison with state-of-the-art methods

We compare our method with several state-of-the-art approaches on three video anomaly detection datasets, shown in Table 1. In comparison to the baseline method [16], our method achieves improvements of 1.3%, 1.8% and 0.8% on the UCSD Ped2, CUHK Avenue and ShanghaiTech datasets, respectively.

Table 1. Performance comparison of the proposed method with SOTA methods in terms of AUC (%) on three datasets.

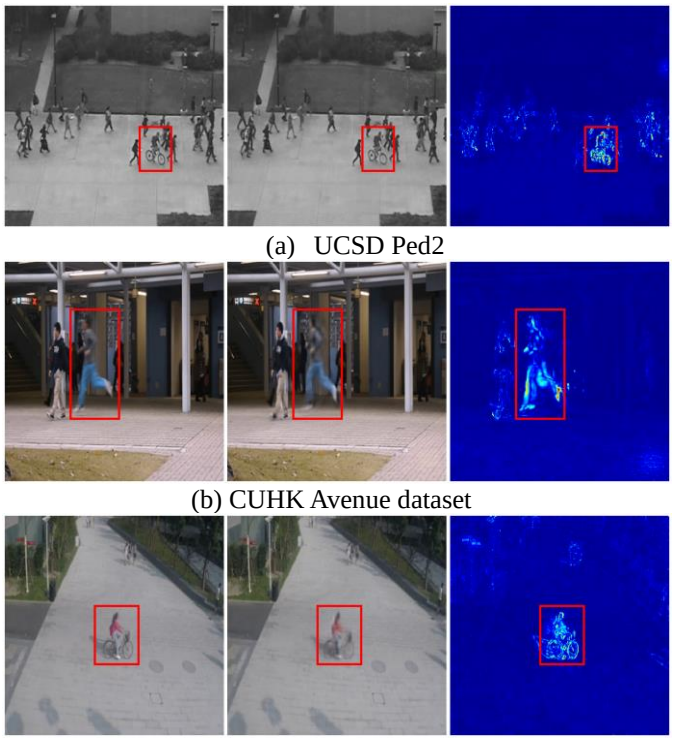
Method	Ped2	Avenue	ShanghaiTech
Luo et al. [1]	96.2	85.7	73.0
Yang et al. [2]	93.7	83.2	
Szymanowicz et al. [3]	84.4	75.3	70.4
Wang et al. [4]	96.2	-	72.5
Saypadith and Onoye et al. [5]	95.7	86.8	73.0
Li et al. [6]	95.4	96.0	71.4
Liu et al [16]	95.4	85.1	72.8
Ours	**96.7**	**86.9**	**73.6**

4.3. Visualization of prediction error

In Figure 2, we visualize several abnormal samples, showing the ground truth in the first column, the predicted frames by our model in the second column, and the prediction error maps in the third column. As shown in Figure 2, the predicted frames of abnormal events are blurrer than the ground truth frames. Consequently, the prediction error maps exhibit significant errors around abnormal objects, such as the cyclist in the first row, the running man in the second row, and the cyclist in the third row.

5. Conclusion

We have presented a transformer-based network for video anomaly detection under unsupervised learning. The network leverages a multi-scale transformer to extract hierarchical features. A convolutional decoder is used to predict the future frame. The proposed method is evaluated on three video anomaly datasets and achieves better performance compared to recent methods.



(a) UCSD Ped2

(b) CUHK Avenue dataset

(c) ShanghaiTech dataset.

Figure 2. Visualization examples of prediction error. From left to right, we show the target frames (first column), predicted frames (second column), prediction error maps (third column). The lighter colors in the error map denote larger prediction error.

References

[1] Luo, Weixin, Wen Liu, Dongze Lian, and Shenghua Gao. "Future frame prediction network for video anomaly detection." IEEE transactions on pattern analysis and machine intelligence 44, no. 11 (2021): 7505-7520.

[2] Yang, Fan, Zhiwen Yu, Liming Chen, Jiaxi Gu, Qingyang Li, and Bin Guo. "Human-machine cooperative video anomaly detection." Proceedings of the ACM on Human-Computer Interaction 4, no. CSCW3 (2021): 1-18.

[3] Szymanowicz, Stanislaw, James Charles, and Roberto Cipolla. "X-MAN: Explaining multiple sources of anomalies in video." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3224-3232. 2021.

[4] Wang, Tian, Xing Xu, Fumin Shen, and Yang Yang. "A cognitive memory-augmented network for visual anomaly detection." IEEE/CAA Journal of Automatica Sinica 8, no. 7 (2021): 1296-1307.

[5] Saypadith, Savath, and Takao Onoye. "Video anomaly detection based on deep generative network." In 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5. IEEE, 2021.

[6] Li, Qun, Rui Yang, Fu Xiao, Bir Bhanu, and Feng Zhang. "Attention-based anomaly detection in multi-view surveillance videos." Knowledge-Based Systems 252 (2022): 109348.

[7] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pvt v2: Improved baselines with pyramid vision transformer." Computational Visual Media 8, no. 3 (2022): 415-424.

[8] Tang, Yao, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. "Integrating prediction and reconstruction for anomaly detection." Pattern Recognition Letters 129 (2020): 123-130.

[9] Liu, Ting, Chengqing Zhang, Xiaodong Niu, and Liming Wang. "Spatio-temporal prediction and reconstruction

network for video anomaly detection." Plos one 17, no. 5 (2022): e0265564.
[10] Yang, JingXian, YiHeng Cai, Dan Liu, and Jin Xie. "3D U-Net for Video Anomaly Detection." In Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering, pp. 1640-1645. 2021.
[11] Hao, Yi, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. "Spatiotemporal consistency-enhanced network for video anomaly detection." Pattern Recognition 121 (2022): 108232.
[12] Le, Viet-Tuan, and Yong-Guk Kim. "Attention-based residual autoencoder for video anomaly detection." Applied Intelligence 53, no. 3 (2023): 3240-3254.
[13] Park, Chaewon, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. "FastAno: Fast anomaly detection via spatio-temporal patch transformation." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2249-2259. 2022.
[14] Chen, Dongyue, Pengtao Wang, Lingyi Yue, Yuxin Zhang, and Tong Jia. "Anomaly detection in surveillance video based on bidirectional prediction." Image and Vision Computing 98 (2020): 103915.
[15] Gong, Dong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1705-1714. 2019.
[16] Liu, Wen, 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, pp. 6536-6545. 2018.
[17] Qiang, Yong, Shumin Fei, and Yiping Jiao. "Anomaly detection based on latent feature training in surveillance scenarios." IEEE Access 9 (2021): 68108-68117.