A GAN-Based Framework Combining Memory and Self-Attention Mechanisms for Video Anomaly Detection in Online Gaming Environments

Li-ting Xiong[1*], Bin Ou[2] and Zhi-Ping Cheng[3]

[1,2,3]School of Artificial Intelligence, Nanchang Jiaotong Institute, Nanchang, Jiangxi, 330110, PR China,
[1]02081@ncjti.edu.cn,[2]BinOU28@outlook.com, [3]Zhiping56@outlook.com

Corresponding author: Li-ting Xiong, 02081@ncjti.edu.cn

Abstract. Traditional generative adversarial networks (GANs) exhibit excessive generalization ability when predicting abnormal samples, which may leads to unstable prediction results. To address this issue, a memory-based GAN is proposed for video anomaly detection tasks in real-world scenarios. Firstly, a memory module is introduced into the adversarial learning framework of the convolutional autoencoder-based prediction network and discriminator network to construct a memory adversarial network that enhances the model's prediction ability for normal video frames. Secondly, a global self-attention mechanism is embedded in the generator, which assigns greater weight to more important information and acquires global features at the same time. A new loss function based on feature compactness and separability is designed for the memory adversarial network to improve the reliability of the training process. Finally, an anomaly evaluation criteria based on memory loss is proposed to enhance the accuracy of anomaly detection. The fusion of PSNR values between future frames and predicted frames and distances between video frame features and memory features further enhances the anomaly detection performance of the model.

Keywords: Video anomaly detection; GAN; Convolutional AutoEncoder; Memory mechanism; Self-Attention; in Online Gaming Environments
DOI: https://doi.org/10.14733/cadaps.2024.S5.91-105

1 INTRODUCTION

In recent years, public safety and stability have become a focus of public attention. Surveillance video, with its features of recordability and analyzability, has played a significant role in maintaining social order by providing valuable clues for law enforcement agencies, and has greatly promoted the

development of smart and safe cities [6],[13]. Consequently, intelligent video surveillance technology has emerged and rapidly become a research hot topic in both academic and industrial fields [18].

With the rapid development of artificial intelligence (AI) technology represented by various deep learning (DL) models, intelligent video surveillance systems based on AI have made significant breakthroughs in pedestrian [10] and vehicle detection [19], face recognition [4], and other fields. However, due to the low quality of cross-scene monitoring videos and the low intelligence level of computer vision algorithms, the degree of intelligence of surveillance systems has not yet reached the needs of practical applications [15]. Effective video anomaly detection technology can accurately detect abnormal behavior in monitored areas and promptly issue corresponding alerts, thus minimizing potential losses in terms of life and property due to unexpected events.It highlights the importance of video anomaly detection in online gaming environments and introduces the GAN-based framework that incorporates memory and self-attention mechanisms.

DL-based anomaly behavior detection can be classified into three categories: supervised learning, weakly supervised learning, and unsupervised learning. Supervised learning methods typically require a large number of labeled samples to train the model [1], while weakly supervised methods can use video-level annotations for learning, and unsupervised methods do not require any labeled data at all [7]. Due to the high cost of data labeling, which often requires extensive human involvement and professional knowledge, it is difficult to collect a large number of labeled samples in practice. In comparison, unlabeled samples are easier to obtain, making unsupervised anomaly behavior detection gradually becoming a research hotspot. Unsupervised methods typically use a certain metric to study the relationship between samples, and then classify and assign unlabeled samples accordingly. As a representative of unsupervised network structures in recent years, generative adversarial network (GAN) have received widespread attention in academia due to their powerful generative capabilities [14]. In the field of anomaly behavior detection, GANs are used to reconstruct or predict video frames, and then detect anomalies based on reconstruction errors, effectively alleviating the problems of underfitting and low detection accuracy caused by insufficient labeled data.

In GAN-based video anomaly detection methods, the interested regions of the input video frames are first detected and features are extracted. Based on these behavior features, abnormal behaviors are detected and classified. The reconstruction-based approaches assume that the reconstruction network only works on normal samples and cannot reconstruct abnormal samples well, so the reconstruction error can be used to distinguish between normal and abnormal samples. Sabokrou et al. [16] Proposed to use the generator to reconstruct video frames while implicitly repairing abnormal regions, and the discriminator was used to judge the possibility of different regions in the video frames being abnormal. The intersection of the outputs of the two networks were deemed as the final anomaly detection result, and the network can locate abnormal behaviors. Zaheer et al. [20] trained the adversarial network for anomaly detection by changing the basic function of the discriminator from distinguishing real and fake data to identifying the quality of reconstructed data. The whole network is continuously optimized through an adversarial feedback loop, and finally generates stable and high-quality data. Due to the indistinguishability of abnormal behaviors, Atghaei et al. [2] proposed to use transfer learning to enable the network to have effective spatiotemporal features and improve algorithm adaptability. Shin et al. [17] utilized the discriminator of GAN as a basic model and applies transfer learning to the anomaly classifier. Since GAN can generate data that does not exist in the dataset, the basic model can learn from data that is similar to real data, solving the problem of insufficient labeled data. The prediction-based anomaly detection methods follow the idea that normal events are predictable while abnormal events are not, and distinguishes between normal and abnormal behavior by comparing the test frames with the predicted ones. Compared to reconstruction-based method, it can break through the limitation of

reconstruction error and increase the difference between normal and abnormal frames. Liu et al. [12] used U-Net as the generator of GAN for prediction, and then estimated the corresponding optical flow. The model was optimized based on the difference between the predicted frame and the original frame, as well as adversarial loss. The cross-layer transfer characteristics of U-Net can effectively retain the basic structural characteristics of the input frame, making the network more focused on the difference between the output and input frames during training. Lee et al. [9] combined GAN and Long Short Term Memory (LSTM) networks, inputting the previous and next five frames of the given frame into the forward Convolutional LSTM (ConvLSTM) and backward ConvLSTM respectively to extract spatiotemporal features. Then, the in-between frame was generated based on the output of LSTM, and the mean square error between the predicted and real frames, as well as the weighted sum of discriminator output, were used as the anomaly score. To address the problems of scale variation and complex motion, feature aggregation network was combined to learn normal patterns at different scales, improving the robustness of detecting complex events [8]. To address the problem of insufficient motion feature extraction, Chen et al. [3] designed a loss function based on the target frame and bidirectional predicted frames, and proposed an abnormal assessment method based on sliding windows to force the network attention focusing on foreground targets of the predicted frame, effectively suppressing noise in the prediction error map, and improving the accuracy and robustness of the detection model.

Video anomaly detection can be regarded as a one-class unsupervised learning problem. Due to the absence of abnormal samples, the generalization performance of the model on abnormal samples is difficult to estimate. If the model has too strong generalization ability, it cannot effectively detect abnormal samples; while if the generalization ability is too weak, it cannot identify new normal samples. To solve the above problems, this paper proposes a video anomaly detection model based on a memory-adversarial network by integrating memory modules with adversarial learning. The specific contributions are as follows:

1) A memory-adversarial network architecture is proposed, which combines a memory module with a generative adversarial network, addresses the issue of over-generalization in video anomaly detection models.
2) An abnormal behavior detection combining global self-attention and convolution operation is proposed, in order to solve the problem that the convolutional neural network (CNN) cannot learn the global interaction between different scenes.
3) A multi-task loss function, comprising of prediction loss, feature reduction loss, and feature diversity loss, is designed to enhance the reliability and accuracy of anomaly detection training.

2 GENERATIVE ADVERSARIAL NETWORK COMBINING MEMORY AND SELF-ATTENTION MECHANISMS

In this paper, a novel memory-adversarial network architecture is proposed by incorporating a memory module into the GAN network based on the frame prediction methodology. The network mainly consists of a generator (prediction network) P, a discriminator network D, and a memory module M. The model is trained only on normal samples. At time t, a sequence of previous normal frames $\boldsymbol{X} = \{\boldsymbol{x}_{t-l},...,\boldsymbol{x}_{t-1}\}$ is fed into P, and the corresponding features $\boldsymbol{q}_t$ are extracted by the encoder. The model retrieves a combined feature $\hat{\boldsymbol{p}}_t$ based on similarity weighting from the memory module, which is concatenated with $\boldsymbol{q}_t$ to obtain feature $\boldsymbol{f}_t = [\boldsymbol{q}_t : \hat{\boldsymbol{p}}_t]$, and the memory module is

updated. Finally, the concatenated feature f_t is fed into the decoder to obtain the prediction result $\hat{x}_t$ for the t-th frame, and the calculation process is as follows:

$$\hat{x}_t = P(X;\theta_M) \tag{1}$$

Where $P(\cdot)$ represents the mapping function corresponding to the prediction network of the generator; θ_M is the parameter of the memory module. Using the previous l frames to predict the current frame is beneficial for the prediction network to learn the timing pattern of normal video clips. The network architecture is shown in Fig. 1.

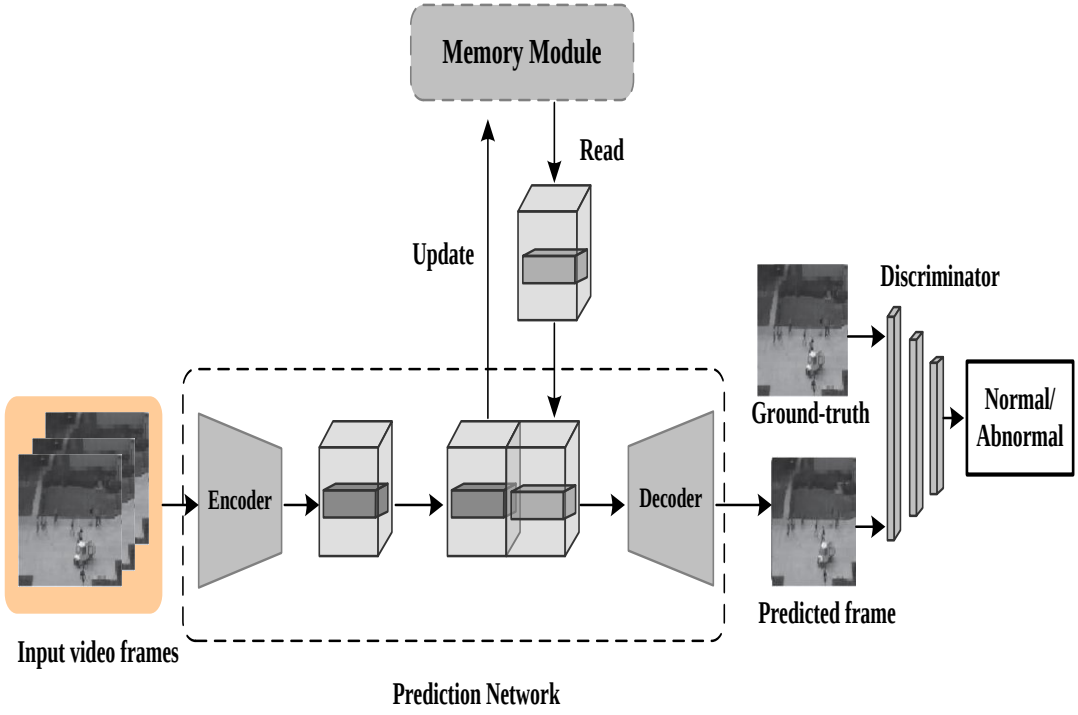


Figure 1: Proposed GAN framework.

In video anomaly detection, the addition of the memory module will further increase the prediction difficulty of abnormal samples, increasing the performance difference of the prediction network on normal and abnormal samples, thus improving the ability of anomaly detection. In the discrimination stage, the predicted samples $\hat{x}_t$ are treated as negative samples, while the ground-truth video frames x_t are treated as positive samples, and both are fed into the the discriminator network D for supervised training. The network trainings are completed by iterative adversarial learning between the prediction network P and the discriminator network D alternatively. In the testing stage,

the prediction error of P and the recognition result of the D are used as the discrimination criteria to obtain an anomaly score.

2.1 Generator Model

The generator of the proposed video abnormal behavior detection framework is shown in Fig. 2. The model is an encoding-decoding structure, and a memory module is added in the middle to store iterative normal behavior features. The model is mainly composed of three modules: AutoEncoder, Bottleneck and Memory.

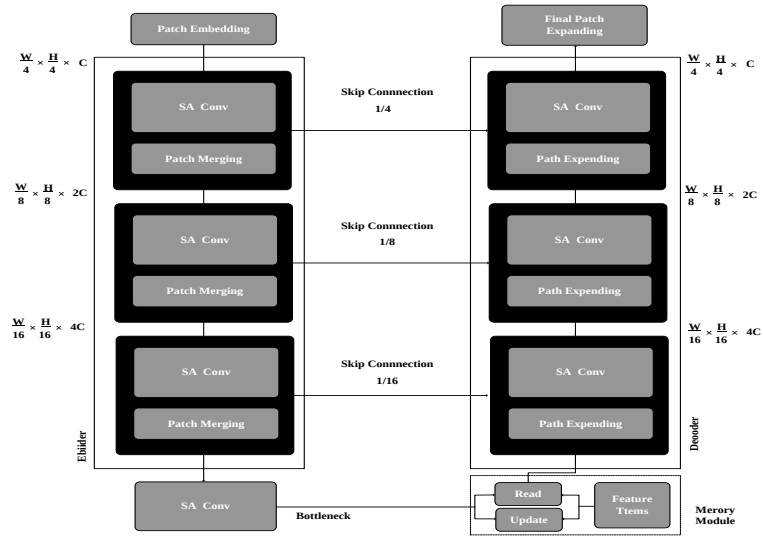


Figure 2: Structure of the Generator Model.

The AutoEncoder module consists of Patch Embedding, self-attention (SA) Conv Block, Patch Merging, and Patch Expand. Patch Embedding converts the matrix form into a sequence form by dividing the image into a series of non-overlapping 4×4 patches, each patch having a feature size of 4×4×3. Subsequently, a convolutional layer is applied to increase the dimensionality of the features to C. The SA Conv Block comprises of Layer Normalization (LN), Windows Multi-head Self-Attention (W-MSA), Shift Window Multi-Head Self-Attention (SW-MSA) [5], Multilayer Perceptron (MLP), BN+ReLU, and Conv. The architecture of the SA Conv Block is illustrated in Fig. 3.

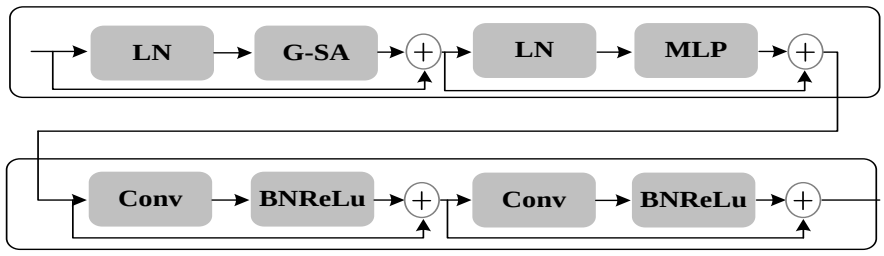


Figure 3: Structure of the SW-Conv model.

The SA Conv Block consists of attention mechanism and convolution. The first layer is a combination of attention mechanism and MLP, and input is passed through a regularization layer. The second layer is composed of convolution, BN normalization, and ReLu activation function. Each module is then connected using skip connections. The global attention in SA Conv Block comes from the attention mechanism, which is essentially a mechanism for reassigning weights to obtain important features from the data. By making the model learn these features in detail, detection performance can be improved. Based on this, a global attention mechanism is realized, which focuses the model attention on the interaction between local and global features, reducing dependence on external information and making it better at capturing internal correlations in the data or features. Compared to the attention mechanism, global attention adds the initial feature map in the feature map fusion, further enhancing the interaction between local and global feature maps.

2.2 Discriminator Model

The discriminator D is trained to distinguish between predicted video frames and ground-truth video frames to achieve adversarial learning with the prediction network. In the proposed framework, a CNN consisting of 5 convolutional layers and 1 fully connected layer is used as the basic architecture for the discriminator. The first 3 convolutional layers use 5x5 convolutional kernels, and the last 2 convolutional layers use 3x3 convolutional kernels, with a stride of 2 for all. ReLU is used as the activation function, and the output channels of the 5 convolutional layers are 64, 128, 256, 512, and 512, respectively. The architecture of the discriminator network is shown in Fig. 4. During training, the ground-truth video frame x_t is used as the positive sample, and the predicted result $\hat{x}_t$ is used as the negative sample, and traditional binary supervised learning is conducted based on cross-entropy loss. The final output of the discriminator is a scalar value between 0 and 1, which represents the discriminator's judgment of the authenticity of the input image. It is used for adversarial learning with the prediction network during the training phase and for identifying abnormal samples during the testing phase.

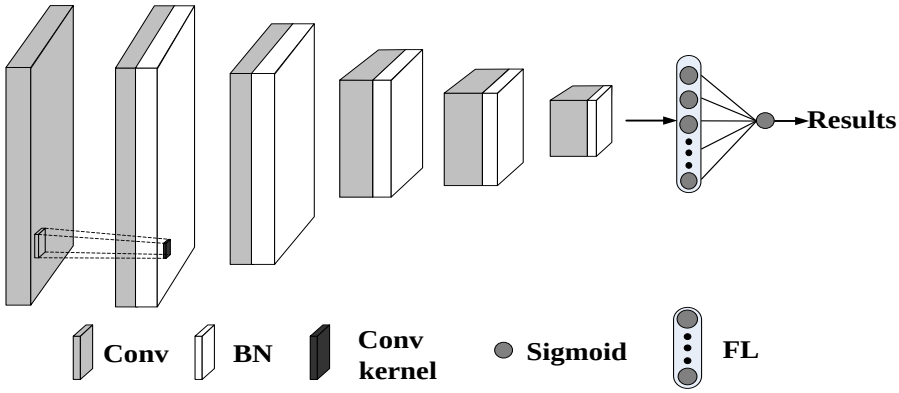


Figure 4: Structure of the Discriminator.

2.3 Global Attention Mechanism

In the proposed attention mechanism, the attention of the model is focused on the interaction between local and global information, reducing reliance on external information and improving its ability to capture internal correlations within data or features. The global attention adds an initial feature map in the feature fusion process, further enhancing the interaction between local and global feature maps, and utilizes linear transformation instead of 1x1 convolutions for feature extraction

during transformation, disrupting the specific structural information obtained from feature maps. Figure 5 illustrates the schematic diagram of the attention mechanism.

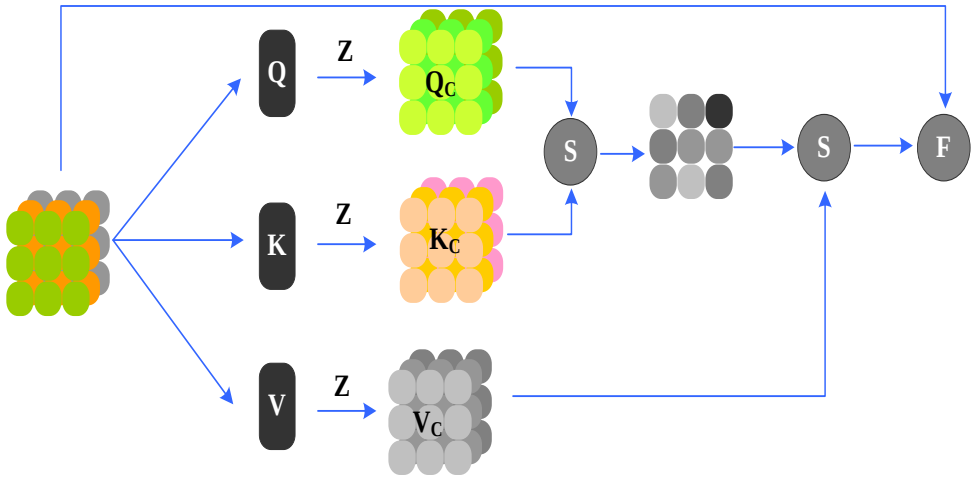


Figure 5: Global Attention Mechanism

First, the obtained features are transformed into three linear feature vectors Q, K, and V. Then, the feature maps are convolved to recover their initial shapes and obtain Qc, Kc, and Vc. The features Qc, Kc, and Vc are then fused sequentially and combined with the original features to obtain the global attention feature map:

$$F_i = Z(Linear_i(\boldsymbol{X})), i \in (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{2}$$

$$GSA(X) = SoftMax(\frac{\boldsymbol{Q}_c \boldsymbol{K}_c^{\mathrm{T}}}{\sqrt{d}})\boldsymbol{V}_c + \boldsymbol{X} \tag{3}$$

The SW-Conv Block module is designed to preserve the dimensions and resolution of the input features so that it can be inserted into any part of the neural network. Since the computational complexity of global attention is higher than that of convolution under the same conditions, the model's computation should be minimized during sampling. This can be achieved by rearranging the features for upsampling and downsampling. Patch Merging downsamples the features, for example, from H/4×W/4×C to H/8×W/8×4C, while Patch Expand is the opposite, for example, from H/8×W/8×4C to H/4×W/4×C. Upsampling and downsampling using Patch Merging or Expand increases the dimension of the features, which significantly reduces the computation burdens compared to convolutional pooling and deconvolution. To avoid the phenomenon of network non-convergence caused by an increase in model depth in the Bottleneck module, the proposed framework uses an SA Conv Block module as the bottleneck of the autoencoder to learn normal features. This allows for the learned features to be represented in a compressed manner.

2.4 Memory Module

In video anomaly detection, it is important to design neural networks that can learn normal sample features effectively while also directing attention to the Memory module as much as possible. This

enables the normal sample features learned at the bottleneck to be preserved. Due to the large amount of normal sample features obtained during model training, it is necessary to reduce them to obtain a limited number of normal sample features. The Memory module includes two operations, read and update. The specific process of read operation is as follows:

The encoder features $\boldsymbol{q}_t$ of size H×W×C are divided along the channel dimension into H×W query items $\boldsymbol{q}_t^k$, each $\boldsymbol{q}_t^k$ with a size of 1×1×C, corresponding to a feature description of a certain spatial position. For each query item $\boldsymbol{q}_t^k$, the cosine similarity $\boldsymbol{w}_t^{k,n}$ with each feature $\boldsymbol{p}_n$ in the memory module is computed using the Softmax function:

$$\boldsymbol{w}_t^{k,n} = \frac{\exp((\boldsymbol{p}_n)^{\mathrm{T}} \boldsymbol{q}_t^k)}{\sum_{n'=1}^{N} \exp((\boldsymbol{p}_{n'})^T \boldsymbol{q}_t^k)} \tag{4}$$

By using the similarity $\boldsymbol{w}_t^{k,n}$ as weighting factor and computing the weighted sum of all features in the memory module, a composite feature $\hat{\boldsymbol{p}}_t^k$ corresponding to $\boldsymbol{q}_t^k$ can be obtained:

$$\hat{\boldsymbol{p}}_t^k = \sum_{n=1}^{N} w_t^{k,n} \boldsymbol{p}_n \tag{5}$$

After each query $\boldsymbol{q}_t^k$ is matched with its corresponding $\hat{\boldsymbol{p}}_t^k$, all $\hat{\boldsymbol{p}}_t^k$ will form a feature tensor $\hat{\boldsymbol{p}}_t$ with the same size of $\boldsymbol{q}_t$. Then, the feature tensor is concatenated with $\boldsymbol{q}_t$ along the channel dimension, and the resulting feature tensor $\boldsymbol{f}_t = [\boldsymbol{q}_t; \hat{\boldsymbol{p}}_t]$ is used as input to the decoding network to generate the prediction result $\hat{\boldsymbol{x}}_t$.

In the memory feature updating process, the matching features are defined as the feature $\boldsymbol{p}_{n\sim}$ in the memory module that have the highest cosine similarity with their corresponding encoding features $\boldsymbol{q}_t^k$. After providing corresponding memory features $\hat{\boldsymbol{p}}_t^k$ for all encoding features $\boldsymbol{q}_t^k$, the memory features $\boldsymbol{p}_n$ are updated by using the matching encoding features $\boldsymbol{q}_t^k$:

$$f(\boldsymbol{p}_n) + \sum_k \frac{\boldsymbol{w}_t^{k,n}}{\max(\boldsymbol{w}_t^{k,n})} \boldsymbol{q}_t^k \tag{6}$$

2.5 Joint Loss Function

To better predict anomalous behavior, two aspects of the loss function should be considered in the model training process. One is the overall network prediction loss function Lrec, the other one is the

loss function of the memory module. Two aspects should be considered in Lm: feature reduction loss Lc and feature diversity loss Ld.

The purpose of Lc is to perform aggregation and dimension reduction of similar normal features. However, this may cause all features in $\boldsymbol{p}_n$ to become more similar with each iteration, which contradicts the initial purpose of this loss function and does not reflect the diversity of normal features. Therefore, Ld is introduced to increase the differences between features in Pn, reduce their similarity, and thereby increase the diversity of normal patterns. The total loss function is calculated using the following formula:

$$L = L_{rec} + \alpha L_c + \beta L_d \tag{7}$$

The prediction loss uses the L2 norm to penalize the difference between the predicted frame $\boldsymbol{x}_t$ and the ground truth frame $\hat{\boldsymbol{x}}_t$:

$$L_{rec} = \sum_{t}^{T} \| \hat{\boldsymbol{x}}_t - \boldsymbol{x}_t \|_2 \tag{8}$$

The feature reduction loss Lc aims to make the input feature more similar to the most similar item in Pn. Therefore, L2 norm is used to penalize the difference between them. Pmax refers to the most similar feature in Pn, which is defined as the value of n when k is fixed and the value of $\boldsymbol{w}_t^{k,n}$ is the maximum:

$$L_c = \sum_{t}^{T} \sum_{k}^{K} \| \boldsymbol{q}_t^k - \boldsymbol{P}_{\max} \|_2 \tag{9}$$

The feature diversity loss Ld aims to encourage Pn to learn diverse normal features. To achieve this, an additional term Psecond is added to Lc to increase the differences between features. Psecond is the second largest value of $\boldsymbol{w}_t^{k,n}$ for a fixed k. In addition, a margin is added to the formula. The calculation formula for Ld is as follows:

$$L_d = \sum_{t}^{T} \sum_{k}^{K} \| q_t^k - P_{\max} \|_2 + \| q_t^k - P_{second} \|_2 + \mu \tag{10}$$

2.6 Anomaly Score

In the proposed memory-adversarial network model, both the consistency between predicted and ground-truth video frames and the similarity between hidden features and memory features can serve as the basis for anomaly evaluation. The former is more sensitive to small-scale spatiotemporal features such as edges, textures, and local motion due to its same image resolution, while the latter is defined in the hidden feature space of the prediction network and is more sensitive to large-scale features such as overall contours, color distribution, and limb motion. Therefore, PSNR and Euclidean

distance are used as quantitative evaluation indicators for anomaly detection, and a combined anomaly score St can be calculated as follows:

$$S_t = \lambda(1 - g(PSNR(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t))) + (1 - \lambda)g(D(q_t, P) \tag{11}$$

where g(·) is a normalization function, which ensures that the normalized PSNR and feature distance scores are in the range of [0,1]. Clearly, a larger St value indicates a higher degree of abnormality.

3 EXPERIMENTS AND DISCUSSIONS

3.1 Datasets

This paper validates the performance of proposed framework using two widely adopted anomaly detection datasets, including UCSD Ped2 [11] and ShanghaiTech Campus [21]. The UCSD Ped2 dataset consists of 16 training videos and 12 testing videos, each containing 180 frames with image size of 240×360. The ShanghaiTech dataset has a larger scale, with a total of 437 videos shot in 13 different scenes, including 330 normal videos and 107 abnormal videos. Each video contains varying numbers of frames ranging from tens to hundreds, and each video frame has a size of 480×856. Normal events involve pedestrians walking normally on the campus, while abnormal events include cars driving on pedestrian paths, violent fights, and robberies.

3.2 Experiment Platform and Evaluation Metrics

The entire experiment was conducted on a Win10 platform with NVIDIA GeForce RTX2060 and Intel(R) Core(TM) i5-10300H CPU @2.5 GHz, using the PyTorch 1.4 deep learning framework. Prior to the experiment, the resolution of the video frames was uniformly adjusted to 224×224, and the pixel values of each frame were normalized to the range of (-1,1). The Adam optimizer was used to optimize the training loss, with an initial learning rate of 1e-4.

Two metrics were used to evaluate the performance of the model. The first metric is the receiver operating characteristic (ROC) curve based on frame-level analysis and the area under the curve (AUC) of the ROC curve. The ROC curve can demonstrate the performance of the classifier by plotting the true positive rate (TPR) and false positive rate (FPR) at different threshold settings. After calculating the above metrics, the ROC curve is plotted, with the FPR on the horizontal axis and the TPR on the vertical axis. The closer the ROC curve is to the upper left corner, the better the performance of the detector. The other metric is the false alarm rate (FA). Since FPR is also known as the false alarm rate, the FPR at a 50% threshold is used as FA. The main part of real-time monitoring videos is normal, so a robust model should also have a low false alarm rate on normal segments.

3.3 Ablation Study

To verify the effectiveness of the improvements of the proposed framework, several combinations of different modules were tested on the UCSD Ped2 and ShanghaiTech datasets. The results are illustrated in Table 1, where Model 1 means that only the prediction loss Lrec was used. It can be observed that adding the memory module and using the discriminator for anomaly detection can effectively improve detection performance on both datasets, and there is a certain complementarity between the two. That is, adding both the memory module and the discriminator can further improve the performance of anomaly detection.

Attention mechanism	*Memory module*	*Joint loss function*	*Discriminator*	*PED2*		*ShanghaiTech*	
				AUC/%	*FA/%*	*AUC/%*	*FA/%*
×	×	×	×	*79.75*	*1.46*	*69.60*	*2.77*
√	×	×	×	*85.54*	*0.75*	*76.53*	*2.30*
√	√	×	×	*94.25*	*0.43*	*82.15*	*1.85*
√	√	√	×	*96.13*	*0.24*	*86.72*	*1.03*
√	√	√	√	*97.75*	*0.12*	*87.85*	*0.76*

Table 1: Results of the ablation study.

A new loss function is proposed for memory-based GAN networks, which introduces two new losses, Lc and Ld, on the basis of the prediction loss Lrec. In order to verify the effectiveness of the proposed loss function in improving the model's anomaly detection performance, a series of ablation experiments were conducted on the ShanghaiTech dataset using different combinations of the two new losses, as shown in Fig. 6. The results showed that after introducing the two new losses of Lc and Ld, the TPR of the model increases from 81.31% to 85.85%, and the inclusion of any of these loss items can improve the performance. This suggested that the proposed multi-loss function improvement is reasonable, complementary, and effective.

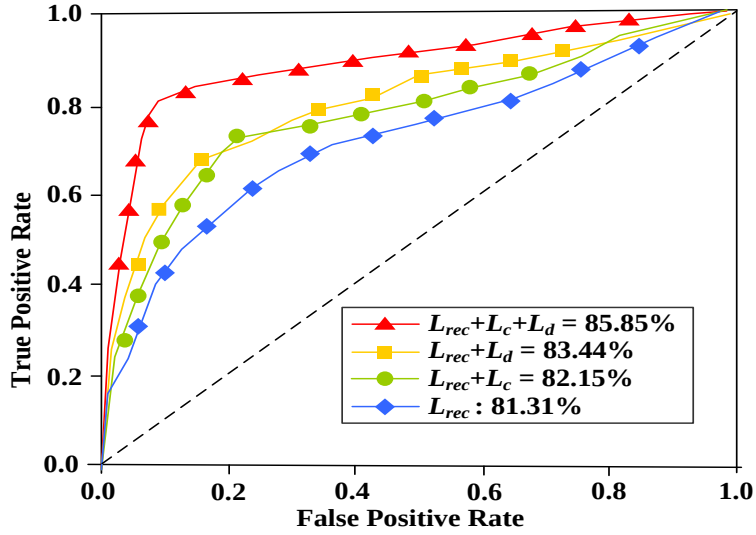


Figure 6: Impact of different loss functions on TPR results.

The selection of weighting coefficients for the discriminator: since the anomaly detection criterion uses the L2 distance in the feature space between input video frames and normal frames (where the memory module records the features of normal video frames) and the weighted sum of video frame prediction errors, the weighting coefficient λ will have a significant impact on the detection results. Therefore, this paper conducted parameter optimization experiments on two datasets to test the performance of the model in detecting anomalies with different values of λ between 0 and 1, in order to find the optimal balance factor λ. The AUC scores of the model on two datasets with different values of λ are shown in Fig. 7. The experimental results showed that the proposed model achieves the best performance on all the datasets when λ=0.6.

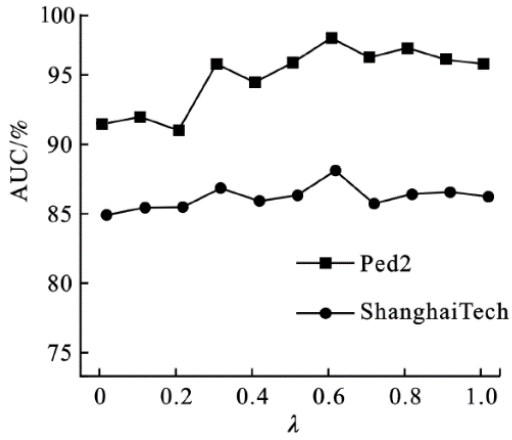


Figure 7: AUC scores with different λ.

3.4 Comparison Experiment

In order to verify the performance of the proposed memory-adversarial network model, we compared the performance of the proposed method with optimal model obtained from the ablation experiments with other anomaly detection methods on the UCSD Ped2 and ShanghaiTech datasets. The results are shown in Table 2.

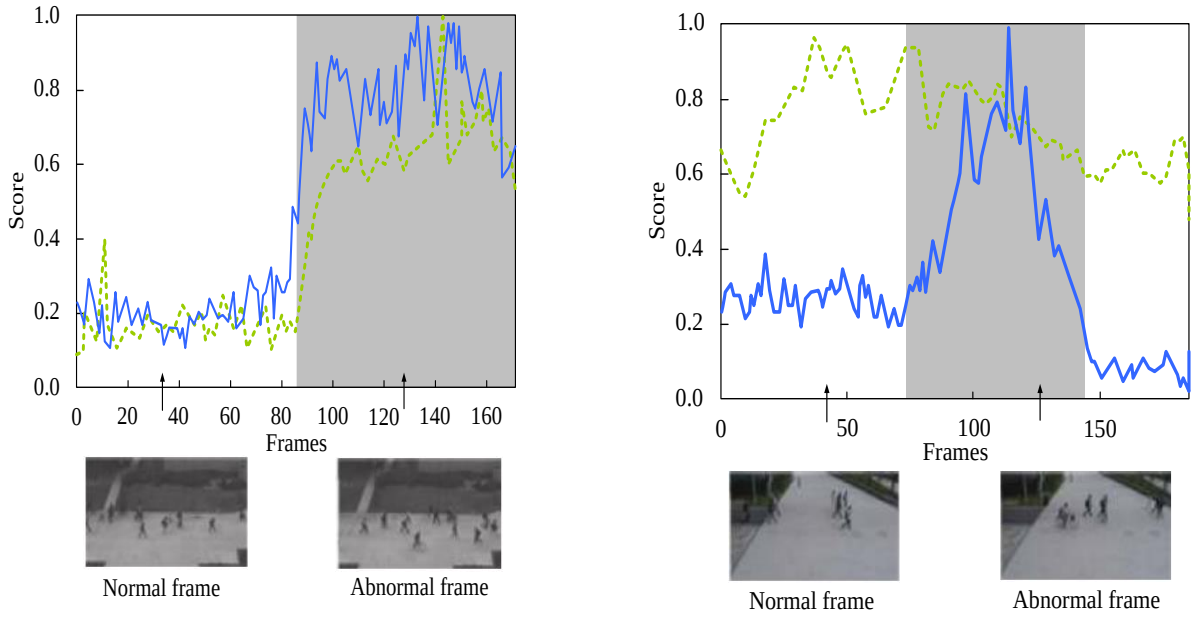
Methods	*Ped2*		*ShanghaiTech*	
	AUC/%	*FA/%*	*AUC/%*	*FA/%*
[11]	*81.74*	*1.23*	*71.55*	*3.22*
[12]	*83.52*	*1.05*	*75.15*	*3.04*
[14]	*86.75*	*0.77*	*77.63*	*2.99*
[15]	*88.15*	*0.62*	*78.45*	*2.71*
[17]	*91.08*	*0.47*	*81.03*	*1.99*
[13]	*95.30*	*0.33*	*85.19*	*1.58*
[18]	*97.01*	*0.24*	*86.34*	*1.35*
Proposed method	*97.75*	*0.12*	*87.85*	*0.76*

Table 2: Comparison results.

It can be seen from Table 2 that compared with the existing unsupervised anomaly detection algorithms, on the UCSD Ped2 dataset, the proposed method achieved the best performance of 97.75% AUC on the 87.85% AUC on the ShanghaiTech dataset. The method in [11] used reconstruction-based detection. Due to the powerful generation ability of GAN and the abnormal events only occupy a small part of the image pixels in the frame, there is no guarantee that there will be a large reconstruction error for the abnormal frame when performing frame reconstruction. The method in [12] introduced a multi-level feature detection method to detect abnormal objects at different semantic feature levels in the video, but its network structure was complex and redundant, and the amount of calculation is large, so it is difficult to achieve real-time detection. Although these reconstruction-based methods can handle abnormal samples, they may also identify new normal samples as abnormal samples. The method in [18] adopted the two-way prediction technique which can effectively extract spatio-temporal features, so it achieved higher detection accuracy than the one-way prediction, but this model can only detect anomalies after the abnormal behavior occurs,

and it is difficult to perform real time detection. The proposed method combined the memory mechanism and the attention mechanism, and achieved the best performance and stability on both datasets, which fully verifies the superiority of the proposed model in video anomaly detection tasks.

The abnormality score curves on representative test videos from two datasets in this study are shown in Fig.8. The solid curve in the figure represents the results with the proposed method, and the dashed curve represents the results with the method in [18]. It can be observed from the figure that when an abnormal event occurred, the abnormality score given by the proposed model increased significantly and quickly dropped to normal level after the abnormal event leaves the monitoring area. The results demonstrate that the proposed model can effectively detect abnormal events in surveillance video data and has accurate and robust detection performance for different types of abnormal events in different scenarios. Moreover, the different scenarios in the two datasets prove that our model not only ensures high abnormal detection accuracy, but also adapts to different scenarios in practical applications, showing strong practicality.



(a) Ped2 dataset (b)ShanghaiTech dataset

Figure 8: Examples of anomaly detection results.

4 CONCLUSION

In this paper, a novel memory-adversarial network model for video anomaly detection is proposed, and the loss function as well as the anomaly scoring method are presented. Through ablation and comparison experiments on two public anomaly detection datasets, we validated the effectiveness and superior performance of the proposed framework. The experimental results demonstrated that the combination of memory module and adversarial network improves the model's ability to learn deep features, thereby enhancing the model's generalization ability to normal samples and discrimination ability to abnormal samples. This conclusion provides a novel and feasible approach for researching abnormal detection in surveillance videos.

Li-ting Xiong, https://orcid.org/0009-0004-0209-0748
Bin Ou, https://orcid.org/0009-0007-5786-8299
Zhi-Ping Cheng, https://orcid.org/0009-0004-8088-9635

REFERENCES

[1] Acsintoae, A.; Florescu, A.; Georgescu, M. I.: et al. Ubnormal: New Benchmark for Supervisedopen-Set Video Anomaly Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 20143-20153. https://doi.org/10.1109/CVPR52688.2022.01951

[2] Atghaei, A.; Ziaeinejad, S.; Rahmati, M.: Abnormal Event Detection in Urban Surveillance Videos Using Gan and Transfer Learning, Arxiv Preprint Arxiv, 2011, 09619, 2020.

[3] Chen, D.; Wang, P.; Yue, L.: et al. Anomaly Detection in Surveillance Video Based on Bidirectional Prediction, Image and Vision Computing, 98, 2020, 103915. https://doi.org/10.1016/j.imavis.2020.103915

[4] Hassan, R. J.; Abdulazeez, A. M.: Deep Learning Convolutional Neural Network for Face Recognition: A Review, International Journal of Science and Business, 5(2), 2021, 114-127.

[5] Jiang, J.; Zhu, J.; Bilal, M.: et al. Masked Swin Transformer Unet for Industrial Anomaly Detection, IEEE Transactions on Industrial Informatics, 19(2), 2022, 2200-2209. https://doi.org/10.1109/TII.2022.3199228

[6] Kashef, M.; Visvizi, A.; Troisi, O.: Smart City as a Smart Service System: Human-Computer Interaction and Smart City Surveillance Systems, Computers in Human Behavior, 124, 2021, 106923. https://doi.org/10.1016/j.chb.2021.106923

[7] Kiran, B. R.; Thomas, D. M.; Parakkal, R.: An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos, Journal of Imaging, 4(2), 2018, 36. https://doi.org/10.3390/jimaging4020036

[8] Lee, S.; Kim, H. G.; Ro, Y. M.: Bman: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection, IEEE Transactions on Image Processing, 29, 2019, 2395-2408. https://doi.org/10.1109/TIP.2019.2948286

[9] Lee, S.; Kim, H. G.; Ro, Y. M.: STAN: Spatio-temporal adversarial networks for Abnormal Event Detection, 2018 IEEE International Conference On Acoustics, Speech And Signal Processing, IEEE, 2018, 1323-1327. https://doi.org/10.1109/ICASSP.2018.8462388

[10] Li, G.; Yang, Y.; Qu, X.: Deep Learning Approaches on Pedestrian Detection in Hazy Weather, IEEE Transactions on Industrial Electronics, 67(10), 2019, 8889-8899. https://doi.org/10.1109/TIE.2019.2945295

[11] Li, W. X.; Vijay, M.; Nuno, V.: et al. Anomaly Detection and Localization in Crowded Scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(1), 2014, 18-32. https://doi.org/10.1109/TPAMI.2013.111

[12] Liu, W.; Luo, W.; Lian, D.: et al. Future Frame Prediction for Anomaly Detection–a New Baseline, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 6536-6545. https://doi.org/10.1109/CVPR.2018.00684

[13] Raaijmakers, S.: Artificial Intelligence for law enforcement: Challenges and Opportunities, IEEE Security & Privacy, 17(5), 2019, 74-77. https://doi.org/10.1109/MSEC.2019.2925649

[14] Ren, J.; Xia, F.; Liu, Y.: et al. Deep Video Anomaly Detection: Opportunities and Challenges,2021 International Conference on Data Mining Workshops, IEEE, 2021, 959-966. https://doi.org/10.1109/ICDMW53433.2021.00125

[15] Rezaee, K.; Rezakhani, S. M.; Khosravi, M. R.: et al. A Survey on Deep Learning-Based Real-Time Crowd Anomaly Detection for secure Distributed Video Surveillance, Personal and Ubiquitous Computing, 2021, 1-17. https://doi.org/10.1007/s00779-021-01586-5

[16] Sabokrou, M.; Khalooei, M.; Fathy, M.: et al. Adversarially Learned One-Class Classifier for Novelty Detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 3379-3388. https://doi.org/10.1109/CVPR.2018.00356
[17] Shin, W.; Cho, S. B.: CCTV Image Sequence Generation and Modeling Method for Video Anomaly Detection Using Generative Adversarial Network,Intelligent Data Engineering and Automated Learning–Ideal 2018: 19th International Conference, Madrid, Spain, November 21–23, 2018, Proceedings, Part I 19. Springer International Publishing, 2018, 457-467. https://doi.org/10.1007/978-3-030-03493-1_48
[18] Sreenu, G.; Durai, S.: Intelligent Video Surveillance: a Review Through Deep Learning Techniques for Crowd Analysis, Journal of Big Data, 6(1), 2019, 1-27. https://doi.org/10.1186/s40537-019-0212-5
[19] Wang, H.; Yu, Y.; Cai, Y.: et al. A Comparative Study of State-of-the-Art Deep Learning Algorithms for Vehicle Detection, IEEE Intelligent Transportation Systems Magazine, 11(2), 2019, 82-95. https://doi.org/10.1109/MITS.2019.2903518
[20] Zaheer, M. Z.; Lee, J.; Astrid, M.: et al. Old is gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 14183-14193.
[21] Zhang, Y.; Zhou, D; Chen, S.: et al. Single-Image Crowd Counting Via Multi-Column Convolutional Neural Network, Las Vegas: IEEE Conference on Computer Vision and Pattern Recognition, 2016. https://doi.org/10.1109/CVPR.2016.70