Video Sequence Anomaly Detection with Multi-Layer Memory-Augmented Autoencoder

Minxiang Long[1], Pengwei Zhang[1], Jingxia Chen[1,*], Wentao Lin[1], Yiyi Gao[1]

[1]College of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, 710021, China
**Corresponding author*

Abstract: *In order to enhance the model's ability to learn normal pattern features in the video anomaly detection task, this paper proposes an end-to-end video anomaly detection method that combines reconstruction and prediction. The method consists of two modules: (1) multi-layer memory-enhanced auto-encoder module: reconstructs RGB frames using a multi-layer auto-encoder with skip connection to compensate for the information loss due to memory; (2) conditional variational auto-encoder module: the reconstructed RGB frames from the previous step are taken as inputs, and predicts future frames using the current optical flow as a condition to capture the correlation between the optical flow and the video frames. Comparative experiments are conducted on Avenue, Ped2, and SHTech datasets, and the experimental results show that the hybrid model achieves relatively strong anomaly detection capability.*

Keywords: *video anomaly detection, Self-encoder, variational inference, memory-augmented*

1. Introduction

Video anomaly detection refers to identifying events in a video sequence that do not fit the expected behavioral pattern[1]. This is an open and challenging research direction because the number of anomalous events is often much smaller than the number of normal events, and the anomalous events themselves are difficult to define accurately in practice[2]. Obviously, pre-collecting and categorizing all abnormal event types is difficult to achieve. Therefore, models used for video anomaly detection tend to be unsupervised, where the model learns the features of normal events and models them, and all events that are categorized outside of the normal pattern are identified as anomalous events.

In the field of video anomaly detection, reconstruction-based and future frame prediction approaches are two common paradigms. Reconstruction-based methods [3, 4] generally use a self-encoder or Unet to reconstruct normal events. Since these models only learn the normal patterns in the training set, abnormal data input to the reconstruction model can lead to large reconstruction errors during the testing phase, thus differentiating it from normal data. Prediction-based methods[5, 6] generally utilize the temporal features of video sequences to train a network and predict the next frame based on the current number of frames, using the prediction error as a metric for anomaly detection.

There are also some works[7, 8]that combine the two paradigms as a hybrid framework for video anomaly detection. Although these methods can detect anomalies in most scenarios, the accuracy is not very impressive. In this paper, we propose an end-to-end hybrid model containing an improved MemAE method and a CVAE module conditional on optical flow for video anomaly detection.

2. Related work

2.1 Memory network

Memory modules in deep neural networks have attracted a lot of attention in the last few years. Literature[9] proposes a differential computational neural network consisting of a backbone network for extracting deep features and an external memory module dedicated to memorizing normal patterns. Literature[4] proposed for the first time the use of Memory Augmented Auto Encoder (MemAE) for abnormal behavior detection for limiting the generative capacity of neural networks. MemAE receives information from the encoder and matches it as a query to the memory slot that is close to it, and later on combines these memory slots in the form of a weighted sum to generate new encoded features for reconstruction by the decoder. Literature[3] proposes a memory network that can update the memory matrix at test time. In this paper, we improve the traditional memory network by adding memory modules directly between the different layers of the codec, and improve the updating of the memory matrix to make the normal patterns of the memory more representative, so as to better distinguish between normal and abnormal patterns.

2.2 Variational autoencoder

Along with the development of discriminative models, generative models are also advancing, most typically GAN[10] and VAE[11]. Among them, VAE is a directed graph model with hidden variables, which can learn the approximate distribution of hidden variables compared with ordinary AE, and then generate new data by sampling. In the generation process, it is assumed that the input is, the hidden variable is , and according to the Bayesian formula there is:

$$P(Z\mid X)=\frac{P(X\mid Z)P(Z)}{\int P(X\mid Z)P(Z)dZ} \tag{1}$$

The conditional probability is an intractable distribution, and literature[11] proposes to incorporate an identification process to approximate the computationally unsolvable prior distribution by learning the distribution and measuring the degree of approximation between the two distributions through dispersion. In order to solve the structured prediction problem, literature[12] proposes CVAE, which consists of an identification network, a conditional prior network, and a generative network, where the output data, observation conditions, and hidden variables are represented, respectively. Literature [13] followed up the work of CVAE by designing a variational Unet network, which achieves better generation results with the character pose as the condition and the picture appearance information as the input. Literature[14] proposed for the first time video anomaly detection using reconstructed optical flow as input and current frame as condition, however, literature[14] did not use an end-to-end model setup, but instead used the preprocessing method based on FlowNet and Resnet's Spatio-Temporal Cubes (STC) designed in literature[15], to pre-process the video sequences for target detection before the start of the training, and then afterwards, the target's frame sequences and optical flow sequences are extracted for training.

Experiments in literature[13] and literature[14] have demonstrated that image sequences with sufficient correlation with video frame sequences can be used as conditions for conditional variational autoencoders to enhance the performance of generative models. In this paper, we propose an end-to-end CVAE prediction module conditional on optical flow, which memorizes normal patterns and achieves anomaly detection by learning the correlation between optical flow and RGB frames.

3. Methodology

As shown in Fig. 1, the end-to-end hybrid model proposed in this paper consists of a memory network and a CVAE module, and the model as a whole is trained on a training set containing only normal patterns, and the weighted sum of the reconstruction error and prediction error will be used as the anomaly score of the video sequences in the testing phase. The details of the hybrid model are specified in later sections.

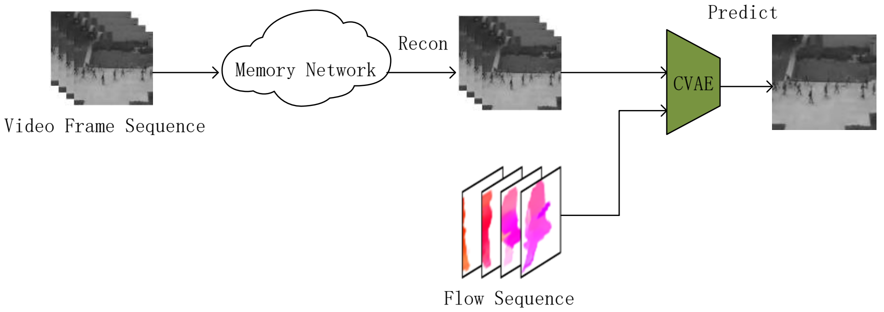


Figure 1 End-to-end video anomaly detection model with integrated reconstruction and prediction modules

3.1 Memory-enhanced autoencoder

The MemAE model aims to solve the problem that self-encoder class models generally have too much generalization ability[4], and consists of an encoder, a decoder, and a memory module. The memory module retrieves the items with high relevance in the memory storage unit through an addressing operation based on the attention mechanism, and then reconstructs the high-dimensional features through these items.

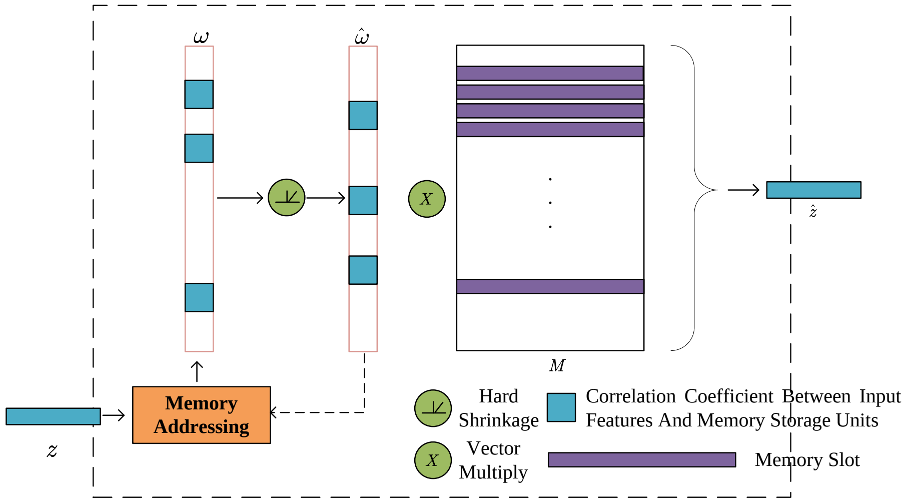


Figure 2 MemAE structure diagram

The structure of the memory-enhanced auto-encoder is shown in Fig. 2, and the model stores the coded features of all normal samples in the memory matrix. ω is the correlation coefficient between the current input coded feature and the memory matrix M, and $\hat{\omega}$ is the correlation coefficient after hard compression. The hard compression forces the model to take fewer memory items to reconstruct the coded features z.

The Memory Enhancement module produces an output by multiplying the weight vector with the memory matrix M:

$$\hat{z} = \omega M = \sum_{i=1}^{N} \omega_i m_i \tag{2}$$

where ω is a row vector with nonnegative elements and sum to 1 ,ω_i is an element of ω . The weight vector is obtained from the input z, and is the number of rows of the memorization matrix is N.

During training and testing, ω is calculated by the degree of approximation of z and M:

$$\omega_i = \frac{\exp\left(d\left(z,m_i\right)\right)}{\sum_{j=1}^{N} \exp\left(d\left(z,m_j\right)\right)} \tag{3}$$

$d(z,m_i)$ represents the proximity measure of feature z and i-th item m_i of the memory matrix m:

$$d\left(z,m_i\right) = \frac{z m_i^T}{\|z\|\|m_i\|} \tag{4}$$

According to Eq $(2)\sim(4)$. The memory enhancement module reconstructs the features $\hat{z}$ by retrieving the entries stored in the memory matrix M that are closest to the features z. In order to suppress the reconstruction of abnormal inputs, the correlation coefficients ω are constrained and the weight coefficients below the sparsity threshold are set to 0 for hard compressed.

$$\hat{\omega}_i = h\left(\omega_i;\lambda\right) = \begin{cases} \omega_i, \omega_i > \lambda \\ 0, otherwise \end{cases} \tag{5}$$

Because ω it is not a continuous function, we cannot do error back-propagation during the training process, and we use a continuous activation function $RELU$ to construct new constraint coefficients:

$$\hat{\omega}_i = \frac{\max\left(\omega_i - \lambda, 0\right)\cdot\omega_i}{\left|\omega_i - \lambda\right| + \epsilon} \tag{6}$$

Here ε is a very small positive number.

3.2 Memory matrix update method

Updating the Memory matrix in the literature[4] is done automatically by the neural network, and this approach causes the distance between different memory slots of Memory to shrink, resulting in model collapse. In this paper, a new way of updating memory slots is proposed, for each memory slot, the query with which the distance is less than τ is selected, and the set of subscripts of the query corresponding to the m-th memory slot is set to be U_t^m, through the following equation:

$$p^m \leftarrow f\left(p^m + \Sigma_{k\in U_t^m} v'^{k,m}_t q_k^t\right) \tag{7}$$

The slot is updated, where f represents the L2 regularization and $v_t^{k,m}$ represents the match of the query to the slot, as follows:

$$v_t^{k,m} = \frac{\exp\left(\left(p_m\right)^T q_t^k\right)}{\Sigma_{k'=1}^{K} \exp\left(\left(p_m\right)^T q_t^{k}{}'\right)} \tag{8}$$

These weights are later renormalized:

$$v'^{k,m}_t = \frac{v_t^{k,m}}{\max_{k\in U_t^m} v_t^{k',m}} \tag{9}$$

The impact of the query closest to the slot can be better represented by using a weighted average sum of these queries, rather than simply adding them together.

For the loss function in the training phase, in order to make the memory slots representing the same query more compact with each other, a compactness loss is introduced in this paper:

$$\zeta_{compact} = \sum_t^T \sum_k^K \| q_t^k - p_p \|_2 \tag{10}$$

At the same time, a separation loss is introduced to prevent the compactness loss from making the representation memory slots homogeneous:

$$\zeta_{sep} = \sum_t^T \sum_k^K \left[\| q_t^k - p_p \|_2 - \| q_t^k - p_n \|_2 + \gamma \right] \tag{11}$$

Accordingly, for the calculation of the anomaly score at the frame level in the testing phase, it is based on the feature compactness index of the interrogation and the nearest memory slot, in addition to the quality of the frame reconstruction:

$$D\left(q_t, p\right) = \frac{1}{K}\sum_k^K \| q_t^k - p_p \|_2 \tag{12}$$

3.3 Multiscale Memory Networks with Jump Connections

Common MemAE approaches typically place a memory module between the encoder and the decoder, but a single memory module does not sufficiently memorize the normal patterns in the video sequence, allowing for better reconstruction of certain abnormal frames as well. To solve this problem, an intuitive solution is to place memory modules between both encoder-decoder corresponding layers, but this in turn leads to excessive feature filtering, i.e., only the most representational features can be memorized, and model collapse may occur.

In this paper, we draw on the idea of Unet to add jump connections between different layers of the encoder and decoder, and add memory modules between the jump connections, and the improved MemAE architecture is shown in Fig. 3. The purpose of this is, on the one hand, to make the raw information from the encoder pass directly to the decoder, which provides more information to the memory modules at different layers for better memorization of the normal patterns. On the other hand, the decoder obtains a higher level of encoded features, which can be more easily decoded by the decoding network for the input, although filtered by the attention mechanism in the MEMORY module. However, when the data from the outermost encoder is connected to the decoder at the outermost level of the input via a jump connection, the encoded information at the outermost level masks the effect of the information at the inner level, which is equivalent to not adding any MEMORY module, resulting in the failure of the model to converge, and thus in this paper we add a jump connection only between the two inner levels.

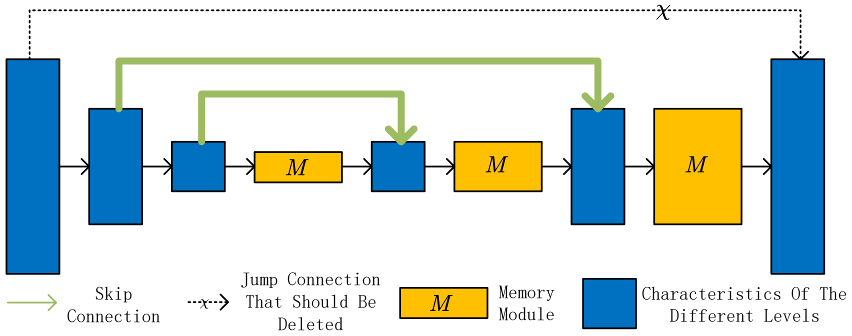

χ
M
M
M
Skip Connection
Jump Connection That Should Be Deleted
M
Memory Module
Characteristics Of The Different Levels

Fig. 3 Structure of a multi-layer MemAE with the addition of a jumper module

3.4 CVAE model for future frame forecasting

Future frame prediction is another common paradigm for video anomaly detection, which has its own characteristics with respect to reconstruction modeling. It is generally modeled $p(x_{t+1}|x_{1:t})$ so that the current t frame $x_{1:t}$ is used to predict the future frame x_{t+1}. Some works[2, 16] have tried to use optical flow information as an aid to help generate RGB frames, but direct fusion of bimodal information is often ineffective. In this paper, we propose a conditional variational self-encoder module for direct modeling $p(x_{t+1}|x_{1:t}, y_{1:t})$, which achieves a better fusion of optical flow modality and video frame module.

According to the theory of variational inference in the literature[13], in order to learn the potential relationship $p(x_{t+1}|y_{1:t})$ between optical flow and RGB frames , the following equation can be obtained:

$$\begin{aligned} \log p(x_{t+1}|y_{1:t}) &= \log\int_z p(x_{t+1},z|y_{1:t})dz \\ &= \log\int_z \frac{p(x_{t+1},z|y_{1:t})}{q(x_{t+1},z|y_{1:t})}q(x_{t+1},z|y_{1:t})dz \\ &\geqslant \mathbb{E}_q \log\frac{p(x_{t+1},z|y_{1:t})}{q(z|x_{t+1},y_{1:t})} \\ &\geqslant \mathbb{E}_q \log\frac{p(x_{t+1}|z,y_{1:t})p(z|y_{1:t})}{q(z|x_{t+1},y_{1:t})} \end{aligned} \tag{13}$$

where the inequality relation in the third line is obtained from Jensen's inequality [17]. Since $x_{1:t}$ and x_{t+1} come from very short intervals in the video sequence, they are very close to each other, for which it can be assumed that they are determined by the same hidden variables z. The above equation can be obtained by replacing $q(z|x_{t+1},y_{1:t})$ to $q(z|x_{1:t},y_{1:t})$:

$$\begin{aligned}\log p &\geqslant \mathbb{E}_q \log \frac{p(x_{t+1}|z,y_{1:t})\,p(z|y_{1:t})}{q(z|x_{t+1},y_{1:t})}\\ &\approx \mathbb{E}_q \log \frac{p(x_{t+1}|z,y_{1:t})\,p(z|y_{1:t})}{q(z|x_{1:t},y_{1:t})}\\ &= -KL\left[q(z|x_{1:t},y_{1:t})\,\|\,p(z|y_{1:t})\right]+\mathbb{E}_q\left[p(x_{t+1}|z,y_{1:t})\right]\end{aligned} \tag{14}$$

Based on the above formulas, this paper proposes a future frame prediction module based on CVAE, as shown in Fig. 4

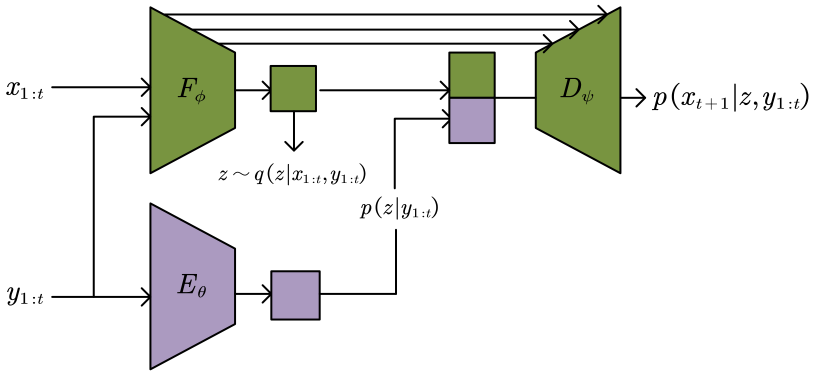


Fig.4 Future frame prediction module based on CVAE

The module contains two encoders E_θ and F_ϕ, a decoder D_ψ,where E_θ encodes the optical flow $y_{1:t}$ into $E_\theta(y_{1:t})$, thus learning the prior distribution $p(z|y_{1:t})$. The inputs F_ϕ are spliced with $x_{1:t}$ and $y_{1:t}$, and the output features $F_\phi(x_{1:t},y_{1:t})$ are learned from the posterior distribution $q(z|x_{1:t},y_{1:t})$. In the training process, the predicted future frames are obtained by sampling z from the posterior distribution $q(z|x_{1:t},y_{1:t})$ and splicing it with the condition $E_\theta(y_{1:t})$ into the decoder D_ψ. In addition, the module adds jump connections between F_ϕ and D_ψ to help predict future frames.

The loss function of a traditional VAE network generally consists of KL scatter and prediction error as defined below:

$$\begin{aligned}\zeta_{VAE} = \|x_{t+1}-\hat{x}_{t+1}\|^2 +\\ KL\left[q(z|x_{1:t},y_{1:t})|p(z|y_{1:t})\right]\end{aligned} \tag{15}$$

In order to make the reconstructed image sharper, this paper also adds the gradient error on top of that:

$$\begin{aligned}\zeta_{gra} = \sum_{i,j}\left(|X_{i,j}-X_{i-1,j}|-|\hat{X}_{i,j}-\hat{X}_{i-1,j}|\right)\\ \left(|X_{i,j}-X_{i,j-1}|-|\hat{X}_{i,j}-\hat{X}_{i,j-1}|\right)\end{aligned} \tag{16}$$

where i,j denotes the spatial coordinates of the pixels in the image. In this way, the loss function of the future frame prediction module in this paper is:

$$\zeta = \lambda_{VAE}\zeta_{VAE} + \lambda_{gra}\zeta_{gra} \tag{17}$$

where λ_{VAE} and λ_{gra} belong to the model hyperparameters.

4. Results and discussion

4.1 Datasets

In order to evaluate the performance of the proposed model and compare it with mainstream video anomaly detection algorithms, this paper chooses to conduct experiments on three publicly available video anomaly detection datasets, i.e., UCSD Ped2[18], CUHK Avenue[19] and ShanghaiTech[20].

UCSD Ped2 consists of 16 training videos and 12 test videos, each with a resolution of 360*240 video frames.The normal data for the training consists of walking pedestrians only, while the abnormal events are due to the circulation of non-pedestrian entities (e.g., automobiles) or abnormal pedestrian movement patterns (e.g., skateboards).

(2) CUHK Avenue consists of 16 training videos and 21 test videos collected from a stationary scenario with a total of 47 anomalous events, e.g., running, bag throwing, etc.

(3) ShanghaiTech is a very challenging dataset in the field of video anomaly detection, it contains videos of 13 scenes with complex lighting conditions and camera angles. The total number of frames for training and testing reaches 274K and 42K, respectively. 130 anomalous events are included in the test set, including chasing and jostling and vehicle approaching.

4.2 Criteria

In this paper, we use the commonly used metrics in the field of video anomaly detection[2, 21, 22] to quantitatively analyze the model performance. By adjusting the threshold of the anomaly score and calculating the true and false cases, we obtain the curve, and use the area under the curve of the frame level to evaluate the model performance. In order to verify the effect of introducing the optical flow information on the video anomaly detection performance, the difference between the average normalized anomaly scores of normal frames and anomalous frames is used to evaluate the discriminative effect of the model on normal and anomalous behaviors, and the larger the difference, the more obvious the discriminative effect of the model is.

4.3 Result and analysis

All the experiments in this paper are done in the configuration of GeForce RTX 3090 GPU from NVIDIA with Intel @ Xeon E5-2603 1.70GHz x6 CPU running on centos, and the deep learning framework chosen is pytorch.

4.3.1 Comparison of results with other algorithms

In order to illustrate the improvement of the model's effect on video anomaly detection, the effect of this paper on three datasets is compared with other existing mainstream methods, and the AUC results of different methods are shown in Table 1.

Table 1 Frame-level AUC values for different methods on Avenue, Ped2, SHTech datasets

Method	UCSD Ped2	CUHK Avenue	SHTech
MemAE[4]	94.1	83.3	71.2
MNAD-P[23]	97.0	88.5	70.5
AMMC[24]	96.6	86.6	73.7
MPN[6]	96.9	89.5	73.8
ABRA[7]	97.4	86.7	73.6
EVAL[25]	97.1	86.7	76.6
TST[8]	95.9	87.2	**78.7**
Our Method	**98.1**	**90.6**	78.6

It can be seen that the performance of this paper's method on Avenue, Ped2, and SHTech datasets are all comparable, especially the SHTech dataset, which contains 13 different road scenarios with different definitions of normal and abnormal events in different scenarios, which increases the difficulty of the model's experiments on this dataset, and the performance enhancement of this paper on the SHTech dataset further proves the enhancement of the model's ability to identify the abnormal events.

Since the training set contains only normal patterns, the anomalous patterns input to the model will show higher anomaly scores than the normal patterns, in Fig. 5, when a video sequence is anomalous (a person running on the sidewalk), the model gives a higher anomaly score to the sequence of frames in which the person runs, and when the anomaly is over (the person runs out of the camera's range) the anomaly scores of the frame sequences return to normal levels.

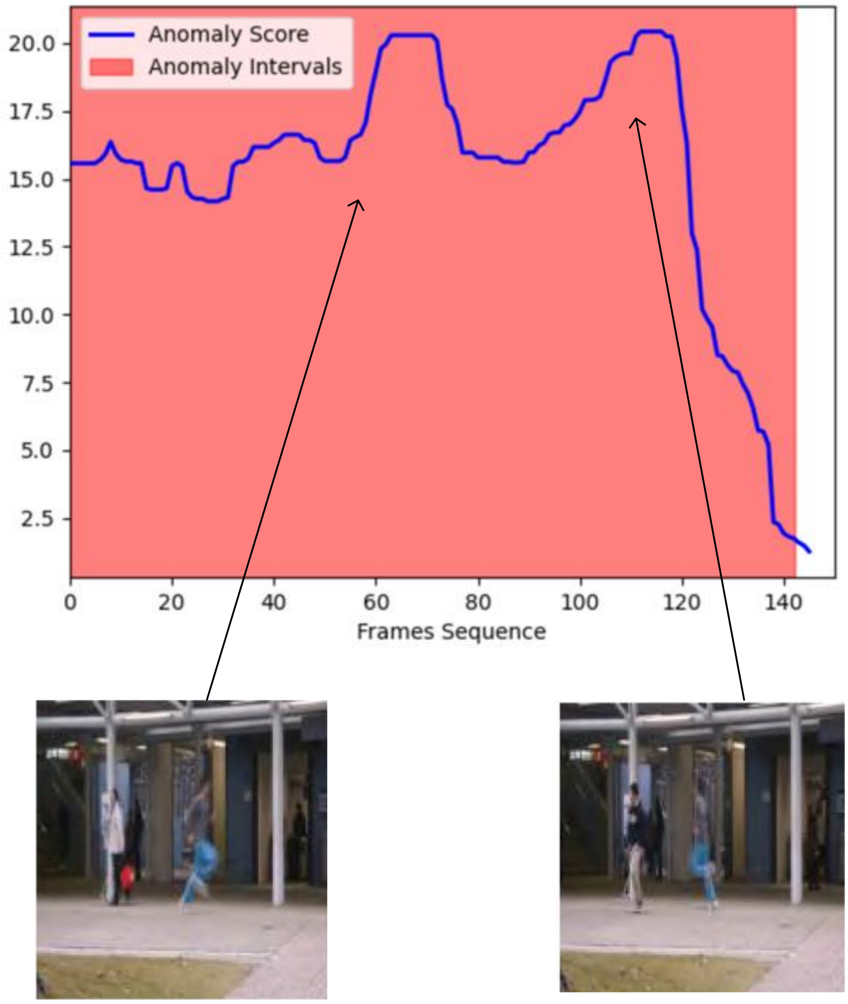


Fig.5 Comparison of anomaly scores in the Avenue test dataset with real scenarios

Overall, both the improvements to MemAE and the addition of the optical flow module effectively enhance the model while enabling end-to-end training, eliminating the need for preprocessing operations with high computational overhead employed in some of the work.

4.3.2 Ablation experiment

1) Impact of adding optical flow information on model performance

In order to test the effect of optical flow modality on this model, this paper compares the effect of predicting future frames using only VAE and predicting future frames using CVAE with the addition of optical flow information, and the results of experiments using only the improved MemAE are used as a control group. To ensure the validity of the experimental results, this paper uses the same initial parameters in training. As can be seen from Table 2, the AUC and ∇s of the model with added optical flow information are higher than those of the improved MemAE model without the optical flow module on the above three datasets, which proves the effectiveness of the module.

Table 2 Comparison of the effect of adding optical flow information or not on model performance

Results	Ped2		Avenue		SHTech	
	∇s	AUC/%	∇s	AUC/%	∇s	AUC/%
MemAE only	0.455	96.7	0.285	84.5	0.173	74.9
VAE	0.443	94.8	0.261	84.1	0.167	74.1
CVAE	0.469	98.1	0.286	90.6	0.181	78.6

By comparing the results of predicting future frames using VAE and CVAE, it can be found that the future frames predicted by directly inputting the RGB frame sequences obtained from MemAE reconstruction into the VAE network are not effective, and even make the accuracy of anomaly detection decrease, this is because the MemAE itself has already made full use of the information of the RGB frame sequences, and the addition of VAE prediction module on the basis of which is not able to uncover more This is because MemAE itself has fully utilized the information of RGB frame sequences, and adding the VAE prediction module on top of it cannot discover more information, so the performance is poor. After the reconstructed RGB frames are taken as input and the optical flow is fed into the CVAE network as a condition, the model utilizes the information of the optical flow modes and the potential relationship $p(x_{t+1}|y_{1:t})$ between the two to make predictions, which results in an obvious improvement of the model performance.

2) Impact of changing the way memory matrix weights are updated

The ablation experiments in this section demonstrate the effect of different loss function and anomaly score calculation methods on MemAE's ability to memorize normal patterns after changing the weight update method, as shown in Table 3.

Table 3 Comparison of the results of tests using different loss functions with the calculation of anomaly scores

Compactness Loss	Seperatness Loss	Characteristic Compactness Index	AUC On the Ped2 Dataset
Y	N	N	94.1
Y	N	N	93.8
N	Y	N	95.5
Y	Y	N	95.5
Y	Y	N	96.2

The first row serves as the reference group and uses only the error between the reconstructed frame and the input frame as the loss function, and the calculation of the anomaly score is based on this error only. The second row adds the compactness loss to the reference group, and it can be seen that the model performance increases rather than decreases, for the same reason as described in Section 3.2, where adding only the compactness loss causes all memory slots to homogenize, the ability to memorize the normal pattern decreases, and the model appears to be underfitted. The fourth row is the result of adding both compactness loss and separability loss, which significantly improves the model performance relative to the first two sets of models, with separability loss making the memory matrix more efficient at memorizing different normal patterns. The third row shows the results of the experiment with only separability loss added, which is not much different from the results of the reference group.

Comparison of the results in the first four rows shows that adding compactness loss alone rather degrades the model performance, adding separation loss alone weakly improves the model, and the combination of the two enables the memory slots of the updated weights of the memoty matrix to achieve a balanced enhancement of the effect in terms of linkage and differentiation of the high relevance queries.

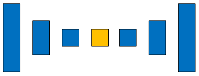
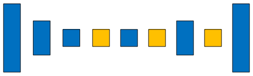
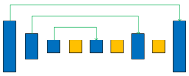
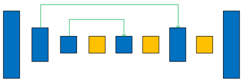
The last two rows show the comparison results of whether or not the feature separation index is added when calculating the anomaly scores, in order to save space and control variables, only the case where both compactness and separation loss are added is used for the comparison experiment here. It can be seen that the group that added the feature separation index when calculating the anomaly scores achieved better model performance, due to the fact that the index measures the differentiation of the different features that represent the same normal pattern, which results in better quantification of the anomalies.

3) Enhancement of Memory Networks by Incorporating Multi-Layer Jump Structures

In order to verify the effect of added memory modules and jump connections on the performance of MemAE models, Table 4 lists the difference between the normalized anomaly scores of normal and anomalous frames and the corresponding AUCs in the Ped2 test set for MemAE models with different added memory modules and jump connections.

In order to verify the effect of added memory modules and jump connections on the performance of MemAE models, Table 4 lists the difference between the normalized anomaly scores of normal and anomalous frames and the corresponding AUCs in the Ped2 test set for MemAE models with different added memory modules and jump connections.

Table 4 Effect of memory modules and jump connections on MemAE performance

	AUC	Δs
	94.1	0.455
	92.3	0.364
	91.2	0.342
	95.4	0.489

5. Conclusion

In this paper, an end-to-end video anomaly detection model is designed based on memory network and conditional variational inference by combining the methods of reconstruction and prediction. Among them, the memory network applies the improved MemAE method, adopts a multilayer structure with jump-joins in the model structure, and improves the updating method of the memory matrix, which can more accurately memorize the normal patterns in the video sequences and produce a large reconstruction error for the anomalous inputs. The conditional variational self-coding model takes the RGB frames reconstructed in the previous step as inputs and conditions the dense optical flow to further enhance the anomaly scores of the frames where the anomalous events are located, thus achieving better anomaly detection. Experiments on common datasets in three domains show that the proposed method outperforms previous reconstruction-only or prediction-only methods, and also outperforms among hybrid methods.

Acknowledgments

This work was financially supported by National Natural Science Foundation of China (61806118) fund.

References

[1] Chandola, V., A. Banerjee, and V. Kumar, Anomaly detection: A survey. ACM computing surveys (CSUR), 2009. 41(3): p. 1-58.
[2] Liu W, Luo W, Lian D, et al. Future frame prediction for anomaly detection–a new baseline[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6536-6545.
[3] Fan Y, Wen G, Li D, et al. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder[J]. Computer Vision and Image Understanding, 2020, 195: 102920.
[4] Gong D, Liu L, Le V, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1705-1714.
[5] Lu Y, Kumar K M, shahabeddin Nabavi S, et al. Future frame prediction using convolutional vrnn for anomaly detection[C]//2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019: 1-8
[6] Lv H, Chen C, Cui Z, et al. Learning normal dynamics in videos with meta prototype network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15425-15434.
[7] Le, V.-T. and Y.-G. Kim, Attention-based residual autoencoder for video anomaly detection. Applied Intelligence, 2023. 53(3): p. 3240-3254.
[8] Cao, C., Y. Lu, and Y. Zhang, Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. arXiv preprint arXiv:2209.02899, 2022.
[9] Graves, A., et al., Hybrid computing using a neural network with dynamic external memory. Nature, 2016. 538(7626): p. 471-476.

[10] Mirza, M. and S. Osindero, Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[11] Kingma, D.P. and M. Welling, Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[12] Sohn, K., H. Lee, and X. Yan, Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 2015. 28.

[13] Esser P, Sutter E, Ommer B. A variational u-net for conditional appearance and shape generation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8857-8866.

[14] Liu Z, Nie Y, Long C, et al. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13588-13597.

[15] Yu G, Wang S, Cai Z, et al. Cloze test helps: Effective video anomaly detection via learning to complete video events[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 583-591.

[16] Nguyen T N, Meunier J. Anomaly detection in video sequence with appearance-motion correspondence[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1273-1283.

[17] Needham, T., A visual explanation of Jensen's inequality. The American mathematical monthly, 1993. 100(8): p. 768-771.

[18] Sabokrou, M., et al., Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, 2018. 172: p. 88-97.

[19] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab[C]//Proceedings of the IEEE international conference on computer vision. 2013: 2720-2727.

[20] Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked rnn framework[C]//Proceedings of the IEEE international conference on computer vision. 2017: 341-349.

[21] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 733-742.

[22] Lai Y, Liu R, Han Y. Video anomaly detection via predictive autoencoder with gradient-based attention[C]//2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020: 1-6.

[23] Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 14372-14381.

[24] Cai R, Zhang H, Liu W, et al. Appearance-motion memory consistency network for video anomaly detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(2): 938-946.

[25] Singh A, Jones M J, Learned-Miller E G. EVAL: Explainable Video Anomaly Localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18717-18726.