Video anomaly detection with NTCN-ML: A novel TCN for multi-instance learning

Wenhao Shao, Ruliang Xiao, Praboda Rajapaksha, Mengzhu Wang, Zhigang Luo, Roberto Minerva, Noel Crespi

Video Anomaly Detection with NTCN-ML: a Novel TCN for Multi-Instance Learning

Wenhao Shao[a,b,*], Ruliang Xiao[c], Praboda Rajapaksha[b], Mengzhu Wang[a], Noel Crespi[b], Zhigang Luo[a], Roberto Minerva[b]

[a] *National University of Defense Technology, College of Computer, 410073, Changsha, China*
[b] *Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France*
[c] *Fujian Normal University, College of Computer and Cyber Security 35007, Fuzhou, China*

Abstract

A key challenge in video anomaly detection is the identification of rare abnormal patterns in the positive instances as they exhibit only a small variation compared to normal patterns, and they are largely biased by the dominant negative instances. To address this issue, we propose a weakly supervised video anomaly detection model called NTCN-ML - Novel Temporal Convolutional Network Multi-Instance Learning Model. The NTCN-ML model extracts temporal representations of video data to construct a time-series pattern to optimize the multi-instance learning process. The model examines the correlation between positive and negative samples in the multi-instance learning process to balance the feature association between rare positive and negative instances. The video anomaly detection with the NTCN-ML model achieved 95.3% and 85.1% accuracy for UCF-Crime and ShanghaiTech datasets, respectively, and outperformed the baseline models.

*Corresponding author:Wenhao Shao, E-mail address:shaowenhao007@gmail.com

Keywords: Video process, Pattern recognition, Anomaly detection, Feature extraction, Temporal convolutional network, Deep Learning

1. Introduction

Video anomaly detection is a significant problem yet an active research area in which models observe patterns that deviate from normal behavior, which serves a crucial role in industrial production and transportation. There are still some challenges and problem complexities that require advanced approaches to model the patterns in complex video data to identify outliers. One main challenge is the recognition of positive instances or rare abnormal patterns as they manifest only small variations compared with normal events. In addition, rare positive instances are largely biased by the dominant negative instances.

In the literature, supervised learning strategies are mostly used for learning abnormal patterns and normal events, which require manually-annotated labels as learning signals [1]. However, in practice, it is challenging to acquire annotated data for all types of anomalous events, and therefore, supervised learning suffers from several disadvantages [2] such as, i) The boundary between normal and abnormal patterns is blurred in many video scenes. Thus, the same event can produce different consequences in different scenes resulting in different classifications. ii) Anomalous video events are featured with temporal properties, but deep learning usually ignores such features in representation learning. iii) Anomalous patterns cover a wide range of situations and it is unrealistic to define all patterns of anomalous events in a single scenario.

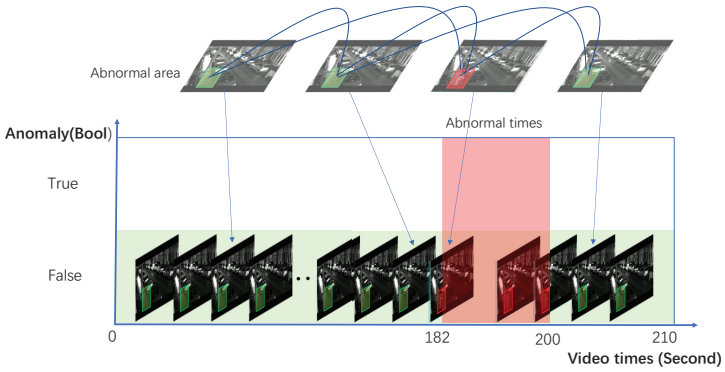


Figure 1: A representation of the spatiotemporal dimension of anomalous events.

To this end, researchers have turned to explore unsupervised learning and weakly supervised learning models for video anomaly detection. Unsupervised methods solely rely on normal events for model training and anomalous events are identified by learning representation features and intrinsic patterns of normal events [3]. Compared to unsupervised algorithms, weakly supervised learning algorithms rely on training samples with both normal and anomalous events. The core of weakly supervised algorithms is the Multi-Instance Learning (MIL)[4]. One assumption in MIL is that the optimization in each training process always targets the negative instance in the abnormal data. However, this assumption is unrealistic as it does not always learn the right patterns, because there is no guarantee that the ranking loss from different scenarios (pairs of normal data and abnormal data) always occurs on the negative instances of abnormal data.

As shown in Figure 2, the error between normal instances and normal instances in abnormal videos is larger than that with abnormal instances, which will cause the model to learn in the wrong direction after training.

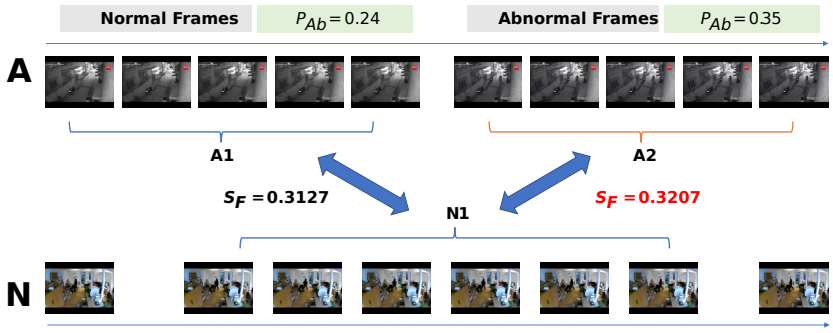


Figure 2: Feature similarity analysis of positive and negative instances: A means abnormal data, N means normal data. $A1$ represents normal instances in abnormal data, $A2$ represents abnormal instances, and $N1$ represents instances in normal data, S_F represents similarity of features between two instances and P_{Ab} is the probability of an anomalous instance.

To mitigate the above issues, this paper proposes to use TCN network to calculate the correlation between positive and negative instances, so as to enhance the temporal characteristics of the model. Inspired by the literature [5][6], which introduced an effective combination of temporal convolution networks and graph neural networks. In this paper, we consider the temporal and spatial features as equally important factors in video anomaly detection and propose a new weakly supervised video anomaly detection model, NTCN-ML (a **N**ew **T**emporal **C**onvolution **N**etwork for **M**ulti-Instance **L**earning). The NTCN-ML model examines the correlation between positive and negative samples in the MIL process to enhance temporal patterns. Positive and negative correlation helps to balance the feature association between posi-

tive and negative instances, and then construct a novel temporal feature to optimize the MIL process.

Two main contributions of this paper are: i) We successfully introduce a novel temporal convolutional network in a weakly supervised learning for video anomaly detection and propose a novel video anomaly detection model NTCN-ML which has optimized the temporal feature extraction and ii) We show that the NTCN-ML model proposed in this paper can effectively learn the potential patterns between anomalous events and normal events. The experimental results on two widely-used benchmark datasets; 1) UCF-Crime dataset - 95.3% accuracy and 2) ShanghaiTech dataset - 85.1% accuracy, show that the performance of NTCN-ML reached state-of-the-art.

2. Background

Existing video anomaly detection models can be mainly divided into three categories: supervised learning, unsupervised learning, weakly supervised learning. Supervised models are limited by data collection, application scenarios, and low scalability[7]. Thus, unsupervised models and weakly supervised models have attracted more attention from researchers.

Unsupervised Learning Video Anomaly Detection

The core of unsupervised learning models is representation learning. Typically, representation learning and self-supervised learning utilize auxiliary tasks to learn valuable features on their own. Future frame prediction and reconstruction[8] are the most common auxiliary tasks. In 2018, Liu Wen et al.[3] proposed an unsupervised video anomaly detection framework (encoder-decoder structure) based on future frame prediction. In 2019, Dong Gong et

al.[9] proposed a deep auto-encoding anomaly detection algorithm for memory storage aggregation, which proposed that due to the excellent representation ability of neural networks, the reconstruction error of anomalous events is not always greater than the threshold. Their proposed model improves the detection of abnormal events. In 2020, Hyunjong Park et al.[10] optimized on the basis of anomaly detection tasks by combining with the U-Net network to further limit the expressive ability of the neural network, and proposed a video anomaly detection algorithm based on future frame prediction and reconstruction. This method saves time and cost and further improves the detection accuracy of abnormal events. In 2021 Zhian Liu et al.[11] proposed HF2-VAD, a hybrid framework that seamlessly integrates stream reconstruction and frame prediction to handle video anomaly detection. Conditional autoencoding and multi-layer memory modules are employed to learn and store the intrinsic patterns of normal events. In 2022, Zaheer et al.[12] proposed a novel unsupervised generative collaborative learning (GCL) method for video anomaly detection, which exploits the low frequency of anomalies to construct cross-supervision between the generator and the discriminator. The method trains two branch networks simultaneously to promote overall convergence. By learning to predict the missing frames of consecutive normal frames, the model can effectively learn various normal patterns in the video.

Weakly Supervised Learning Anomaly Detection

During the training process of many weakly supervised learning video anomaly detection algorithms, vanilla discriminators such as Convolutional 3D (C3D) [13] and Inflated 3D ConvNet (I3D) [14]), are used to extract normal and anomalous samples from the video. A weakly supervised learning

video anomaly detection model was proposed by Waqas Sultani et al. [15] in 2018, which first proposed to use C3D network to extract the video features after clips and input the features into a MIL to calculate anomaly scores for each instance.

In [16], authors proposed a weakly supervised learning scheme with an optimized loss function and adds the in-packet loss of the normal and anomaly packets to the loss function. They initially used an ordinary temporal convolutional network to optimize the input of the MIL. In [17], authors proposed the first weakly supervised learning anomaly detection model for fusion graph convolutional networks, which treats weakly supervised learning as denoising and increases the weight of anomalous instances to improve the reliability of the generated instance labels. Boyang Wan et al.[18] proposed a dynamic MIL scheme, which selects only the k instances with the highest anomaly scores to calculate the anomaly scores and reduces the distance of the instance scores within the normal packet to improve the cohesiveness of normal events, thus improving the performance of the model. Didik Purwanto et al.[19] introduced a temporal relationship network in a weakly supervised learning anomaly detection algorithm to extend features to different scales capturing both short-time dependencies and long-term dependencies. 2022 Huiyu Mu et al.[6] proposed a spatiotemporal mapping convolution-based weak supervised learning anomaly detection model.

However, many previous models ignore the important influence of temporal features on video events. Even though some models try to combine temporal convolution or graph convolution networks to improve the performance of the model, they do not achieve considerable results. Hence, we propose

a weakly supervised learning detection model combining a novel temporal convolutional network, which is divided into two parts. The first part is a temporal convolutional module, which inputs the vanilla features of the video data and outputs a two-dimensional vector indicating its confidence rate of belonging to two categories. Second, in the MIL module, the confidence rate and vanilla features are simultaneously input to the MIL network. A novel weakly supervised learning video anomaly detection model is constructed.

3. Methodology

This section explains a weakly supervised learning video anomaly detection model called NTCN-ML (a New Temporal Convolution Network for Multi-Instance Learning). In general, in the training process of paired data, when the model learns the features of sequential data, the positive and negative instance usually contain a large amount of similar content. Taking the premise that there is a large amount of similar content between positive and negative instance. In the negative samples, the spatio-temporal region where the anomalous events occur only accounts for a very small portion of the entire video, as exemplified by the 23rd video in the Vandalism subcategory in the UCF-Crime dataset [15]. As shown in Figure 1, measured in the time dimension (x-axis), the data unit Vandalism 23 for example, the video lasts about 210 seconds, but the time of the anomalous event occurs lasts only 18 seconds. It is about 8.6% of the whole video. Second, compared to the spatial dimension, the region where the anomaly occurs occupies only a very small number of pixels of the video frame. Figure 1 illustrates that the distinguishing features of abnormal instances in negative samples are not distinctly

prominent. Consequently, achieving accurate optimization of abnormal instances during training becomes challenging. As a result, the optimization of MIL may be steered in the wrong direction. Therefore, it is crucial to enhance the discriminative characteristics of positive and negative instances in the feature space by calculating the correlation between normal instances and abnormal instances in negative samples.

3.1. Temporal Convolutional Networks

1. Design principles

Temporal Convolutional Networks are derived from Time Series Networks. Time series learning networks usually need to follow two principles [20]: (1) The input and output structures of the network should be the same; (2) The features of the current time node are not disturbed by the features of the next time node. The former is used to ensure that in the process of information mining, the sequence feature information will not be reduced and guarantee to extract high-quality representation features. The latter is to comply with objective facts. In the training process when using sequence data, since the complete sequence data has been obtained, the learning model can access the features after the current time node without obstacles. During the application process, the sequence data located after the current node cannot be accessed. Therefore, when designing the learning network structure, we should proceed from practical problems. That is, during the training phase, only the current node is provided with the features of its previous time nodes.

2. Feasibility of retrofitting traditional temporal convolutional networks

In the traditional TCN(Temporal Convolutional Networks) structure, the convolutional network serves as the basic structural unit for extracting temporal features, and there is no aggregation mechanism or large memory module. The traditional TCN model has one-dimensional full convolutional structure[21], and the full convolutional structure ensures that the newly introduced network structure follows the first principle of temporal convolution, i.e., each hidden layer has the same length as the input layer and only the same input and output lengths are satisfied. However, this structure cannot store valid antecedent information and the posterior information may negatively affect the current features in the full convolutional network structure. Therefore, a novel TCN conforming to the second principle is proposed by Cheng et al[22], which consists of a fully convolutional network and a cascaded network. ($TCN = 1DFCN + CausalConvolutions$). The structure of this network implemented using cascading convolution, which uses the features of the same position of the previous layer and the features of its previous position to calculate the features of the current position. This temporal convolutional network conforms to the second principle that the features of the current time node are not disturbed by the features of the next time node, and the model provides stronger theoretical support when dealing with sequential data, such as text data, and video data.

However, this structure also has a major disadvantage, where the sequence of a video data is long and the features of the next layer are being calculated. If all the information before the node in the previous layer is considered as a calculation factor, which will need a very deep network or a very large filter. There is a possibility of parameter explosion, and information that is

too old can also negatively affect the information of the current world nodes and reduce the quality of the extracted features. The existing video detection models usually use graph convolution or LSTM to store the sequence features (temporal features) of video data to complete the detection[23]. It mainly obtains indirect temporal features by LSTM and graph convolution. There is no strict definition and learning of temporal feature information of sequences. Since the convolutional network has a greater ability to scale[24], the performance of convolutional networks is improving in the learning task of sequence models. Based on this, this work introduces the dilated cascade technique into modern convolutional networks and implements a novel temporal convolutional network.

3 The Proposed Novel Temporal Convolutional Network

The proposed TCN consists of Dilated Causal Convolutions (DCC) and residual networks, and the cascaded dilated convolution layer is shown in Figure 3.a. The DCC model was previously mainly used in the field of NLP to increase the perceptual field of view and reduce the computational effort by setting the dilation rate. The residual block network is composed of a series of residual blocks, which is often used to solve the problem of decreasing feature information and increasing training loss caused by the network depth [25]. Its core function can be defined as $X_{T+1} = X_T + F(X_T)$, where X_T denotes the current feature value and F denotes the cascaded convolution function.

Previous studies [26] have shown that TCN models outperform general-purpose recurrent architectures such as LSTM and GRU, and shown that the "infinite memory" advantage of RNN is basically non-existent in practice.

Compared to recurrent architectures, TCN exhibits longer memory and wider convolutional horizons. In recurrent convolutional networks, many advanced schemes for regularizing and optimizing LSTMs have been proposed [27]. These schemes significantly improve the accuracy achieved by LSTM-based architectures on certain datasets. However, in the past two years, before the introduction of architectural elements such as dilated convolution and residual connections, the performance of convolutional architectures did not meet the needs of applications. Simple convolutional architectures are more effective than recurrent architectures such as LSTMs in various sequence modeling tasks. Due to the considerable clarity and simplicity of TCNs, convolutional networks should be seen as a natural starting point and a powerful toolkit for sequence modeling. Video data has sequence properties. In theory, any sequence data can be used to extract temporal features using the TCN model. The proposed TCN network [28] provides an important technique for mining the feature information of video sequences.

3.2. Extraction of Temporal Features of Video Sequences

The proposed TCN network is used to extract the features of the video sequence. This process is mainly divided into three steps: 1) Train vanilla discriminator C3D or I3D to extract the action features of the video data; 2) Input the features extracted by the vanilla discriminator into the new TCN network to extract high-quality temporal features , the steps of this process refer to Figure 3.b. The TCN network introduced in this paper ensures the extraction of high-quality features through multi-layer concatenation and single-layer convolution; 3) According to the final temporal characteristics of the video, set the activation function to identify the

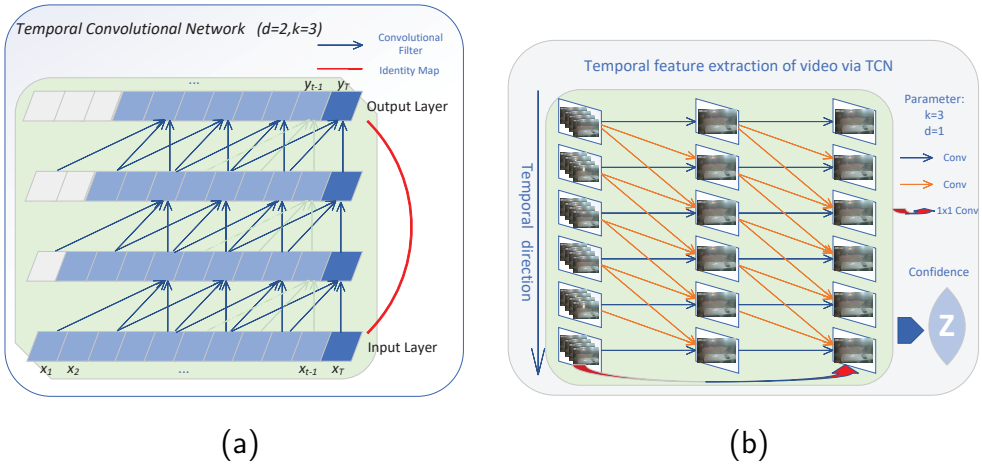


Figure 3: The structure of TCN and Application, (a) The proposed Temporal Convolutional Network structure, under $d = 2$, $k = 3$, the input is $X = x_1, x_2, x_3 \ldots, x_T$, $k = 3$ means that three upper-layer neurons map a neuron of the current layer, $d = 2$ means that the step size; (b) The novel TCN application in video processing. The red line represents the feature of the current node to be extracted, the blue line represents the feature of the previous node, and the red curve represents a $1x1$ convolution unit that retains the most original features of the current node, and the output of the TCN is z, which is the probability value that the input node is an abnormal instance.

video, and calculate the confidence of normal events and abnormal events.

The formalization process and the Qualitative Analysis of the NTCN-ML network:

Consider a video δ is divided into multiple segments δ_i^C, where $(i \in 0, 1, 2, 3 \ldots I)$. The features extracted from the $C3D$ network is represented by: $X_i = \phi_{vanilla}(\delta_i^C)$. Consider all video clip features belonging to the same data unit as a sequence of data $X_1, X_2, X_3, X_4, ..., X_I$, where I represents the number of clips used, the first layer of the hidden layer of the TCN is represented as X^1, and the sequence is represented as $X^1 = X_i^1 | i = 1, 2, 3, \ldots, I$,

the calculation process:

$$X_i^1 = F(\prod_{t=0}^{t=k-1}(X_{i-td})) \tag{1}$$

as shown in Figure 3.b, when $k = 3, d = 1$; then $\delta_i^1 = F(\delta_i \cdot \delta_(i-1) \cdot \delta_(i-2))$, where F represents the convolution function, k represents the kernel during mapping The number, d represents the step size, that is, the distance between two kernel units and so on for the rest of the nodes. The final output of the network structure of the output unit:

$$Output = Activate[(\delta_1, \delta_2, \ldots, \delta_I) + F(\delta_1, \delta_2, \ldots, \delta_I)] \tag{2}$$

Since the video is only divided into normal events and abnormal events, we set the output unit to two node, namely $Output = Z(z_1, z_2)$. z_1 represents the probability that the video unit belongs to normal video, and z_2 represents the probability that the video unit belongs to abnormal video. If the normal video contains elements in some abnormal events, the value of z_1 is more On the contrary, if the abnormal video contains a large number of normal video elements, the value of z_2 is low. Use the formula $\hat{X} = max(z_1, z_2) \cdot X_i$ to construct a new video sequence feature. For normal video δ_n, it belongs to the probability of a positive sample is$max(z_1^n, z_2^n)$, and for abnormal video δ_a its probability belonging to a negative sample is $max(z_1^a, z_2^a)$, then the new feature of normal video $\hat{X}^n = max(z_1^n, z_2^n) \cdot X^n$ the new feature of abnormal video is expressed as $\hat{X}^n = max(z_1^a, z_2^a) \cdot X^a$. The principle it follows is that for any input video feature, only the possibility of $max(z_1, z_2)$ belongs to its true category, and the probability of $min(z_1, z_2)$ will cause misjudgment. There are many factors that cause misjudgment, such as feature entangle-

ment between positive and negative samples, the limitation of neural network learning ability, etc. Excluding uncontrollable factors (a perfect neural network does not exist in practical applications), this work enhances the ability to determine abnormality by improving the separation characteristics between positive and negative samples. Therefore, this work proposes to use disentanglement to improve the performance of instance learning. The process of MIL is a paired training process in which a normal video sample and an abnormal video sample are included, and the probabilities of the normal video and abnormal video belonging to positive and negative samples are different. We have experimented with a variety of new feature calculation methods, and the currently proposed feature calculation method shows a stronger detection performance (here, z_2^a is not directly combined with negative samples, or z_1^n is combined with positive samples, considering that the TCN also exists in the case of misjudgment, we cannot obtain its specific label in advance during the test).

3.3. The NTCN-ML Based on Temporal Convolutional Network Guidance

3.3.1. The proposed NTCN-ML model

A weakly supervised video anomaly detection model based on temporal convolutional network guidance is proposed in this paper. The model uses a novel temporal convolutional network to extract the temporal features of video data and calculates the confidence of the samples. The overall framework is shown in Figure 3. The model also uses the classic vanilla discriminator (C3D - Convolutional 3D, I3D - Inflated 3D ConvNet) to extract the features of the video and combines the obtained confidence with C3D or I3D features to form new input features. Then through the MIL network,

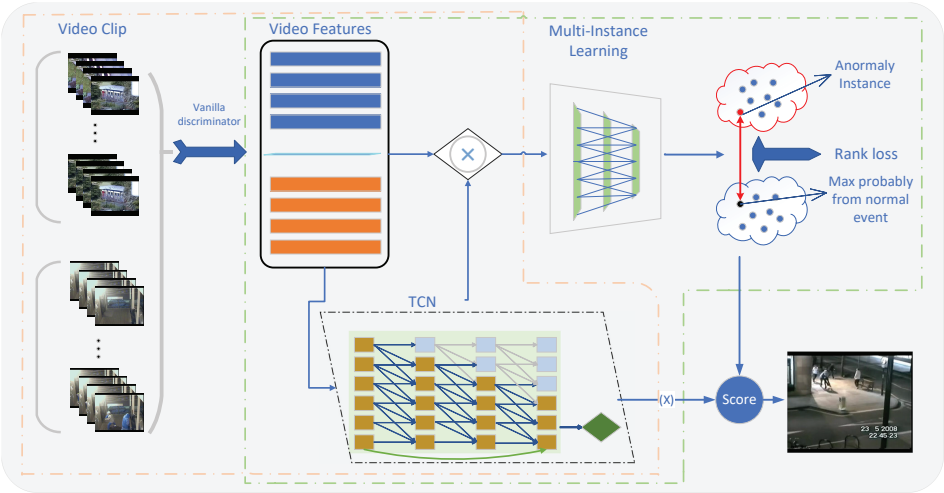


Figure 4: The NTCN-ML framework: The model training process is divided into two phases. The first phase is composed of a vanilla discriminator and novel TCN. The training purpose of this phase is to extract temporal features; the second phase is composed of a vanilla discriminator and TCN. The MIL training is composed of the MIL network, and the novel TCN, this stage is to improve the classification ability of the MIL network.

the final abnormal probability of each instance is calculated; according to the abnormal probability, a loss function is constructed to train the parameters of the MIL. At the same time, the confidence of the video is also involved in the calculation of the abnormal score during the testing process. The NTCN-ML model proposed extracts temporal features through a novel TCN model and enhances the ability of MIL to learn instance labels. Compared with the mainstream algorithms, the NTCN-ML model has a more scientific and effective consideration of temporal features and has stronger robustness. Figure 3 shows the data processing flow of the proposed NTCN-ML model. We discuss model training, loss function, model testing, and algorithm complexity analysis during operation in the following sections.

3.3.2. The Training Phase

The training process is divided into two parts. One is to train the temporal convolutional network. The second is to train a MIL network. The function of the testing phase is to calculate the anomaly score of each instance in the video and locate the time area where the anomaly occurs.

Training the temporal convolutional network is divided into three steps, 1. Input the video clips into the vanilla discriminator to extract features; 2. Input the extracted features into the designed temporal convolutional network (usually select more than 32 clips); 3. Output A 2D array predicting video classification. This two-dimensional array represents the probability that the video belongs to normal events and abnormal events.

The formalization process is as follows: X represents a video, which is divided into multiple segments $X_i, (i \in 0, 1, 2, 3...I)$. Each video segment is called an instance, because sets 16 frames is a segment, so $I = F_n/16$, F_n is the total number of frames in the video. χ is extracts the features of the video segment X_i^C by vanilla discriminator $\phi_{vanilla}$ vanilla is belongs $I3C, C3D$. The TCN function is denoted by f_{TCN}. The final output is represented as:

$$z = f_{TCN} \sum_{i=0}^{I} \chi = f_{TCN} \sum_{i=0}^{I} \phi_{vanilla}(X_i) \tag{3}$$

z represents a two-dimensional vector, where z_1 represents the probability that the video belongs to a normal video, z_2 represents the probability that the video belongs to an abnormal video, the label of the video is $\hbar$, the label of normal video is 0, and the label of abnormal video is 1, If the video is a normal video, its label is $(0, 1)$, otherwise it is $(1, 0)$ at the phase of TCN

training. So the loss of TCN is:

$$loss_{TCN} = Z - \hbar, \qquad \begin{cases} \text{normal} & \hbar = (0,1) \\ \text{abnormal} & \hbar = (1,0) \end{cases} \tag{4}$$

The training of the multi-instance anomaly detection algorithms is divided into four steps. 1. Use the vanilla discriminator to extract video features; 2. The extracted features are input into the pre-trained temporal convolutional network and output a two-dimensional tensor vector; 3. The inner product of the large value and the video feature matrix constructs new video features; 4. The new video features are input into the MIL network, and the abnormal probability of each instance is calculated.

The formalization process is as follows: randomly select the extracted C3D feature $\chi = \phi_{vanilla}(X_i)$ of a fixed length T (fixed number of instances) and the largest value dot product in the output z of the TCN trained in the first phase. Get new video features:

$$\hat{\chi} = z \cdot \chi = f_{TCN}(\sum_{i=0}^{I} \phi_{vanilla}(X_i)) \cdot \chi \tag{5}$$

During the learning process of the multi-instance algorithm, normal videos and abnormal videos are input to the neural network in pairs. We use χ^n n to denote the features of normal videos, χ^a to denote the features of abnormal videos and the MIL is denoted as f_{MIL}.

$$Y = F_{MIL}(\hat{\chi^n}, \hat{\chi^a}) \tag{6}$$

Where $Y = (Y_a, Y_n)$, $Y_a = (y_1^a, y_2^a, y_3^a, ...y_T^a)$ represents the abnormal probability of all instances in the abnormal video package, $Y_n = (y_1^n, y_2^n, y_3^n, ...y_T^n)$

represents the abnormal probability of all instances in the normal video package.

3.3.3. Loss function

The loss function of MIL consists of four parts: ranking loss $L_{ranking}$, smooth loss L_{smooth}, sparse distribution loss $L_{sparsity}$, aggregation loss $L_cluster$. The ranking loss represents the difference between the highest abnormal probability in the normal video package and the highest abnormal probability in the abnormal video package at the training process. So the Ranking loss function is expressed as:

$$L_{Ranking} = ||max(Y_a) - max(Y_n)|| \tag{7}$$

The video is composed of multiple video clips and is sequence data. Therefore, the distribution of abnormal probability should be smooth, and the smooth loss indicates that the occurrence of abnormality in the video sequence is promoted by a process. The smooth loss function is expressed as:

$$L_{Smooth} = \lambda_1 \sum_{i=0}^{T-1} ||y_{i+1}^a - y_i^a||^2 \tag{8}$$

Loss of sparse distribution. In abnormal video, the time of abnormality only accounts for a very small part of the entire video data, so the average abnormal probability of the entire abnormal video is slightly higher than the average abnormal probability of normal video. The sparse loss function is expressed as:

$$L_{sparsity} = \lambda_2 \sum_{i=0}^{T} ||y_i^a - y_i^n||^2 \tag{9}$$

Aggregation loss: The difference between the maximum and minimum values of each instance in the video packets of normal events is not much different. On the contrary, the difference between the maximum value and the minimum value of each instance in the video package of the abnormal event is relatively large. So the aggregation loss is expressed as:

$$L_{cluster} = \lambda_3(1 + max(Y_n) + min(Y_a) - max(Y_a) - min(Y_n)) \tag{10}$$

The total loss function is expressed as:

$$L = L_{ranking} + L_{smooth} + L_{sparsity} + L_{cluster} \tag{11}$$

3.4. The Anomaly Detection Phase

The anomaly detection phase is to describe the detection process of video data that cannot obtain any labels during the test process. The whole process is carried out unsupervisedly. In the detection stage, the algorithm complexity of video anomaly detection is also an important indicator for evaluating models.

3.4.1. Steps of detection

The steps in the anomaly detection phase are divided into five steps. The first step: preprocess the video data, divide the video into multiple video segments, and use the vanilla discriminator to obtain the feature (C3D, I3D) of these segments; The second step: input the feature into the trained

TCN model, obtain the temporal feature and calculate confident level; The third step is to compare the confidence value with the previous one. Feature combination to construct new video features; The fourth step: the video features are input into the MIL, and the anomaly probability of each instance is calculated. In the fifth step, the abnormal probability of each instance is combined with the confidence of the video to obtain the abnormal score. The calculation of the anomaly score, the anomaly score is composed of the last instance anomaly probability, loss, and confidence.

$$Score = z \cdot y + \gamma_1(\delta_y) \tag{12}$$

The pseudocode of the anomaly detection phase is presented by Algorithmic 1:

Algorithm 1 Anomaly Detection

1: Initialization:$f_{TCN}, f_{MIL}, \phi_{vanilla}$, Pre-trained TCN network, MIL network and C3D vanilla discriminator;

2: $\chi = \phi_{vanilla}(X)$, Extract features from video clip X;

3: $Z = f_{TCN}(\chi)$, Calculate the confidence of the video X Equation (3);

4: $Y = f_{MIL}(Z \cdot \chi)$, Calculate the anomaly probability of labels for each segment of the video Equation (6);

5: $Y_{var} = variance(Y)$,Calculate the volatility of video anomaly probability; Equation (12)

Output: Anomaly score=$\{\lambda_1 Y_{var} + Z \cdot Y\}$,

Calculate anomaly scores.

3.4.2. Algorithm complexity analysis in the detection process

The training phase only happens before the model is deployed, so only the algorithmic complexity of the detection process needs to be considered:

The complexity of the temporal convolutional network model: for a video sequence, extract T segments, input the TCN model to classify the video sequence, the algorithm complexity depends on the number of input segments T, the dimension of the feature F of each segment d, the number of hidden layer nodes, the number of hidden layers L, the number of kernels k in the TCN model, the stride d_s, and finally the category C. First, map the extracted features F to the first hidden layer, and each k feature is mapped to a unit.

$$O_{TCN} = (O(\phi_{vanilla}) \cdot k \cdot T)^L \cdot C \tag{13}$$

where C is a 2-category, normal or abnormal, and T is the number of segments, which usually also refers to the number of hidden layer nodes. According to past experience, 32 or 64 are usually chosen, so the algorithm complexity mainly depends on the level of the network and the number of nodes. The complexity of the MIL model: In the MIL process, the input unit usually consists of a feature sequence δ^n from normal videos and a feature sequence δ^a from abnormal videos. Each feature sequence contains T feature segments, and the MIL usually consists of three fully convolutional layers (l_1, l_2, l_3)

$$O_{MIL} = O(F_d(\delta^n) + F_d(\delta^a)) \cdot (l_1 + l_2 + l_3) = O(2F_d \cdot T) \sum_{i=0}^{3} l_i \tag{14}$$

Therefore, in actual operation, the total algorithm complexity is:

$$O = (O(\phi_{vanilla}) \cdot k \cdot T)^L \cdot C + O(2F_d \cdot T) \sum_{i=0}^{3} l_i \tag{15}$$

The above formula shows that the algorithm complexity mainly depends on the number of hidden layers of the neural network and the number of nodes in each layer. This result provides guidance for the design of temporal convolutional network models.

4. Experiments

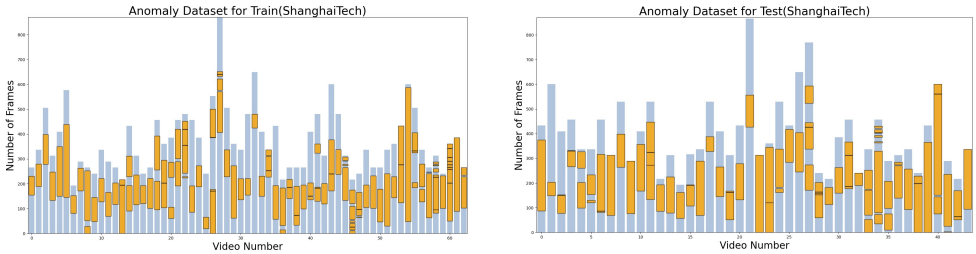
4.1. Datasets

There are two commonly used datasets for weakly supervised video anomaly detection algorithms, namely UCF-Crime and ShanghaiTech datasets, which also is the benchmark datasets. So we validated the proposed model with these two datasets. Table 1 displays the data distribution of the two datasets, revealing that despite the UCF dataset containing an equal amount of normal and abnormal data, the distribution of training and testing sets is imbalanced. Furthermore, during the data reading process, intentional disruption of the sorting of normal data is implemented to enable meaningful comparison and learning between abnormal data and a larger set of normal data

UCF-Crime [15]: is a large-scale dataset consisting of the most primitive surveillance videos. It has a total of 1810 videos, about 200G. Contains 13 common real-world anomalies including abuse, arrest, arson, assault, accident, burglary, explosion, fight, robbery, shooting, theft, shoplifting, and vandalism. The videos are divided into two parts, a training set consisting of 800 normal videos and 810 abnormal videos.

Table 1: Dataset Overview: Nor and Abnor are normal and abnormal videos; Atype is the number of abnormal types; N/A denotes the number of normal videos / abnormal videos

Numbers	Total	Nor	Abnor	ATypes	Train(N/A)	Test(N/A)
UCF_Crime	1700	950	950	13	810/800	140/150
ShanghaiTech	437	330	107	13	175/63	155/44



(a) (b)

Figure 5: The distribution of the ShanghaiTech dataset, (a) denotes the abnormal distribution of abnormal data entries(63 videos) in training, x-axis represents the video number and y-axis represents the total number of frames. Yellow colour is the location for abnormal frames; (b) denotes the abnormal distribution of abnormal data entries in testing for 44 videos.

ShanghaiTech [3]: is a medium-sized dataset of 437 videos with an average of 726 frames per video. The dataset, collected and published by ShanghaiTech University, contains 130 anomalous events in13 scenarios. To make it suitable for evaluating weakly supervised binary classification methods, Zhong et al.[17] split the data into two subsets: a training set consisting of 175 normal videos and 63 abnormal videos. As shown in Figure 5:

4.2. Experiment Details

The core elements in our implementation process include the following steps: 1. When extracting video clips, take 32 video clips equally spaced

Table 2: The TCN classification performance analysis under the UCF-Crime

UCF-Crime(C3D)	32	64	96
32*8	85.1	85.3	84.9
64*8	86	85.1	85.2
128*8	84.8	84.3	84.4

from a video sequence as an example; 2. The extracted C3D and I3D features are stored in numpy format to speed up the training; 3. During the implementation of the TCN network, the core size is still 7, and the number of input channels is 32 for the number of instances of a video. In order to prevent overfitting, The hidden layer of the TCN network used in our work is (32*8), a total of 8 layers, and each layer has 32 nodes; 4. The MIL consists of (512, 32, 1) three-layer fully connected convolutional networks. The evaluation index refers to the literature of this series, with AUC as the main evaluation index[29].

4.3. Experimental results

The experiments are set in three groups. The first group examines the classification performance of the TCN network and obtains the TCN network structure with the best performance. The second group is the AUC evaluation experiment for video anomaly detection. The purpose of this experiment is to measure the performance of the model proposed. The last group is the visualization experiment, the main purpose of which is to promote the important evaluation method for the model to transfer from the experimental scene to the application scene.

The experimental results indicate that with the increase in the number of divided segments, the classification accuracy of TCN does not show a linear

increase. Through the analysis of the dataset, it is inferred that this is due to the "invalid filling" caused by the different time lengths of each video in the data. The reason for data padding is that the duration of some videos is too short to meet the number of divisions, and it is necessary to repeatedly borrow some video clips and video frames to construct a specified number of clips. As an example, 32 video clips are divided into 16 frames, and the total number of video frames of each video cannot be less than 512. Through the analysis of the data set, only a few videos have a total number of frames less than 512. But if the video clip is 64, there are nearly 12% of the data cannot fit into 1024 frames, and therefore overfitting occurs and the detection accuracy decreases. Hence, we decided to use 32 fragments as a reasonable number in our experiments.

4.3.1. **Experiment 2***: AUC comparison with state-of-the-art models*

The purpose of this experiment 2 is to compare the AUC accuracy of the algorithm proposed with the current mainstream algorithms.

In this process, we first train a novel temporal convolutional network. The output value is the probability that the input video belongs to normal video and abnormal video. After completing the TCN model training, input the feature to MIL model. First, divide a video into multiple segments and extract features; second, extract features of a fixed number of segments as input, and the number of segments is the number of input channels; third, input features to the TCN model, calculate the probability; The fourth step is to take the larger value in the two-dimensional array and perform the point multiplication operation with the extracted features, and then input it into the MIL model to calculate the abnormal probability of each segment.

Table 3: Accuracy test of current mainstream algorithms on the UCF-Crime dataset

Method	Source	Technique	Performance (AUC)
Sultani et al[15]	CVPR18	C3D	75.41
TAEDM[30]	SCN20	ResNet	78.51
TCN-IBL[16]	ICIP19	TCN & IBL	78.66
Zaheer et al[31]	SPL21	Self-Reasoning	79.54
GCN-AD.[17]	CVPR19	GCN & Action Classifier	82.12
XD-Violence[32]	ECCV20	Holistic-Localized Networks	82.44
CLAWS[33]	ECCV20	Clustering	83.03
SACRF[19]	ICCV21	Relation-Aware	85.00
RTFM[34]	ICCV21	Feature Magnitude	84.03
STGCNs[6]	IPM22	Spatio-temporal GCN	84.2
BN-SVP[35]	CVPR22	Bayesian	83.39
Ours		Novel TCN	85.1

Table 4: Accuracy test of current mainstream algorithms on the ShanghaiTech dataset

Method	Source	Technique	Performance (AUC)
TCN-IBL[16]	ICIP19	TCN & IBL	83.5
Zaheer et al.[31]	SPL21	Self-Reasoning	84.16
GCN-AD[17]	CVPR19	GCN & Action Classifier	84.44
CLAWS[33]	ECCV20	Clustering-Based	89.67
AR-Net[18]	ICME20	AR Network	91.24
TAEDM[30]	SCN20	ResNet	94.2
MIST[36]	CVPR21	Self-Guided Attention	94.83
BN-SVP[35]	CVPR22	Bayesian	96.0
Ours		Novel TCN	95.3

Table 3 results show that the first list shows that the algorithm proposed in this paper has achieved an accuracy of 85.1% on the I3D features of the UCF-Crime dataset, which has reached the most advanced accuracy. In addition, in order to test the classification performance of the TCN network, extracts the C3D features from the original video to analyze the performance of the TCN, see Experiment 1 for details.

Table 4 shows that the AUC accuracy of the model proposed has reached 95.3%. Compared with the current most mainstream algorithms, the algorithm proposed has surpassed the performance of most published mainstream algorithms. Through the experimental results of the two data sets, it is concluded that the correlation between normal data and abnormal data is also an important consideration in the process of abnormal detection. The model proposed overcomes the above two shortcomings.

4.3.2. **Experiment 3***: Ablation Study*

To test the model's capability, we conducted two sets of ablation experiments: An ablation study and a Loss Function Study. The former involved training and testing different components of the TCN model independently to confirm their effectiveness. The latter involved combining various loss functions during training to examine their impact on performance. Our aim was to verify the impact of different loss functions on the model's performance.

The Ablation study conducted in this paper involves the verification of the model with two datasets (UCF-Crime and ShanghaiTech) using C3D and I3D to independently extract video features and input them into MIL training. Additionally, The model training is divided into four groups, namely I3D+MIL, C3D+MIL, I3D+TCN+MIL, and C3D+TCN+MIL, and the per-

Table 5: Ablation study: Divided two datasets into four groups: I3D+MIL, C3D+MIL, I3D+TCN+MIL, C3D+TCN+MIL, to evaluate the TCN module.

	I3D	C3D	TCN	AUC
ShanghaiTech	✓		✓	0.953
		✓	✓	0.883
	✓			0.861
		✓		0.853
UCF_Crime	✓		✓	0.851
		✓	✓	0.782
	✓			0.823
		✓		0.761

formance was calculated for each group as shown in Table 5.

Table 5 shows the results of the ablation experiments. The results show that the TCN module used in this paper can effectively improve the accuracy of the model on the benchmark. On the ShanghaiTech datasets, compared with the benchmark I3D+MIL, the model proposed in this paper has improved by 9%. Compared with the benchmark C3D+MIL, the model has improved by 3%; For the UCF-Crime dataset, compared with the benchmark I3D+MIL, the model has increased by 3%, and compared with the benchmark C3D+MIL, the accuracy rate has increased by 2%. It shows that the TCN module proposed in this paper is effective

In the study of loss function, this paper uses the ranking loss function as the benchmark, and cooperates with several other novel loss functions to test the performance of the model in two data.

Table 6 shows that when there are more types of loss functions combined, the performance tends to increase slowly. Among them, the $L_{sparsity}$ loss has

Table 6: The Study of Loss Function: Set different loss function combination modes to explore the impact of different loss functions

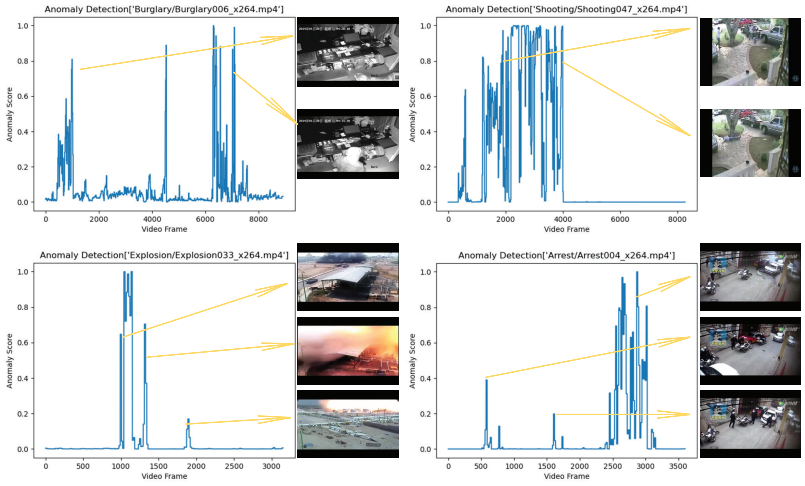
I3D+TCN+MIL	$L_{ranking}$	$L_{sparsity}$	L_{smooth}	$L_{cluster}$	AUC
UCF-Crime	✓				0.836
	✓	✓			0.837
	✓	✓	✓		0.847
	✓	✓	✓	✓	0.851
ShanghaiTech	✓				0.911
	✓	✓			0.913
	✓	✓	✓		0.927
	✓	✓	✓	✓	0.953

a general effect on improving the model performance, and the loss L_{smooth} and $L_{cluster}$ have a greater impact on performance. big. The experimental results show that the l cluster loss function is helpful for performance improvement.

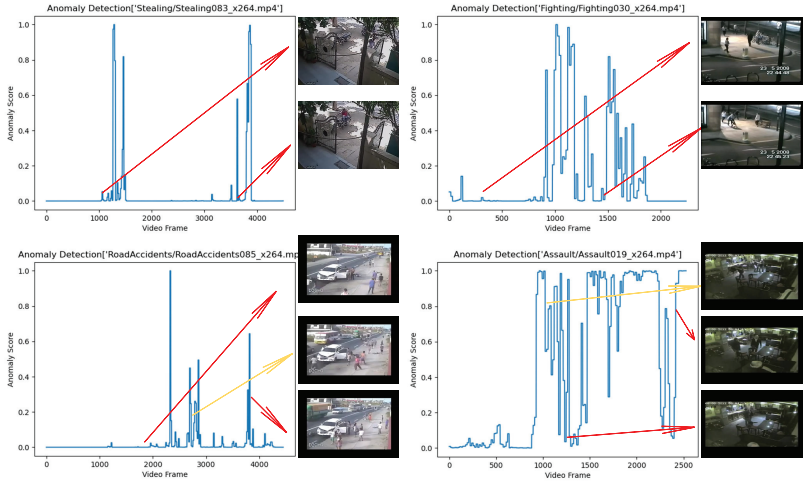
4.3.3. **Experiment 4***: Visual display during anomaly detection.*

In the testing phase, outliers for anomaly detection are constructed from the output of the pretrained TCN network, the output of the MIL model, and the loss function.

The results of experiment 4 show that when an abnormal event occurs in a video, the abnormal score will fluctuate violently (the abnormal score is generated after normalization), so the fluctuation of the abnormal value can be used as the identifier of an abnormal occurrence. Second, the results in figure 6 (b) show that for long-lasting abnormal events, the fluctuations of outliers will be abnormal, resulting in inaccurate detection results. The reasons for this problem mainly come from two aspects: 1. C3D and I3D motion capturers tend to capture short-term actions; the action extractor is



(a)



(b)

Figure 6: Visual effects of the anomaly detection phase, (a) the detection results of anomalous events with a short duration, (b) the detection results of anomalous events with a longer duration d, the yellow line indicates the correctly detected samples, and the red line indicates the detection results is wrong.

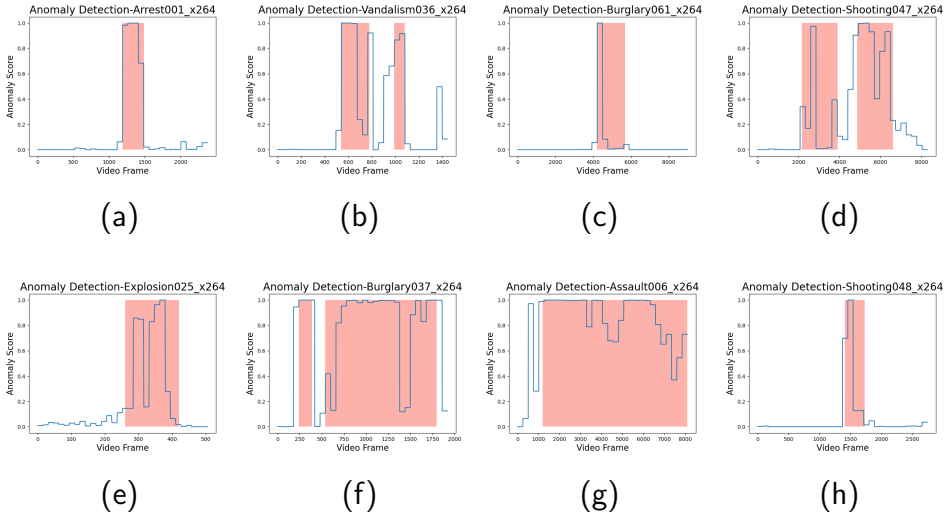


Figure 7: Visual effects of the anomaly detection phase. Red is the area where real anomalies occur, and the curve is the anomaly score.

training in Sports1M, and the action duration of this data set is relatively short. Therefore, the action capture used to preprocess the data set is more favorable for the short duration. 2. Video instance division and the generation of instance outliers do not meet the actual situation of long-term actions. During the experiment, 16 frames are usually delineated as an instance, and there are also cases where the duration is shorter. We hope that follow-up research in this paper can optimize this problem. Overall, the visual effect of this paper is better than other models. Figure 7 is a supplement to the visual experiment. In order to show the experimental effect more clearly, the video data is divided into 32 instances for calculation.

5. Conclusion and Future Work

This work proposes a novel weakly supervised anomaly detection model (NTCN-ML), a new Temporal Convolutional Network (TCN). The NTCN-ML model shows an excellent performance in temporal information mining and provides high-level temporal feature information for weakly supervised learning. The advantage of the NTCN-ML model is that it can enhance temporal features for the entire video sequence, which is different from other related works as they calculate temporal features in segments, and redefine the integrity and coherence of temporal features of video data. Our experimental results show that the NTCN-ML model learns the potential patterns from both anomalous and normal events, and outperformed the baseline anomaly detection models considered in this work. The algorithm presented in this research paper introduces a novel approach for video anomaly detection algorithms, delving into the distribution of data within the feature space in weakly supervised algorithms, and optimizing the process of weakly supervised learning. Furthermore, the proposed model can be seamlessly integrated into other systems, enhancing the algorithm's robustness in real-world applications. However, this paper is subject to certain interpretability limitations. It is expected that future research in the domain of video anomaly detection will primarily focus on improving interpretability. The application of relational triples in NLP represents a promising approach to provide semantic explanations of anomalous events, which is therefore expected to become a primary area of investigation.

In the future, we will asses whether the temporal features extracted from the video sequences align with real-world scenarios, and how the integrity

and coherence of temporal features affect video data analysis. We will also evaluate whether the integrity of temporal signatures has positive implications with both unsupervised and supervised models. Based on the current work, we will further try to define a new anomaly definition in which anomalous events are deeply associated with global temporal signatures. This will certainly help to integrate temporal video analysis patterns in real traffic scenarios.

References

[1] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, L. Shao, Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7842–7851.

[2] B. Ramachandra, M. Jones, R. R. Vatsavai, A survey of single-scene video anomaly detection, IEEE transactions on pattern analysis and machine intelligence (2020).

[3] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection–a new baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6536–6545.

[4] S. Huang, Z. Liu, W. Jin, Y. Mu, Bag dissimilarity regularized multi-instance learning, Pattern Recognition 126 (2022) 108583.

[5] A. Zhao, J. Dong, J. Li, L. Qi, H. Zhou, Associated spatio-temporal capsule network for gait recognition, Vol. 24, IEEE, 2021, pp. 846–860.

[6] H. Mu, R. Sun, M. Wang, Z. Chen, Spatio-temporal graph-based cnns for anomaly detection in weakly-labeled videos, Information Processing & Management 59 (4) (2022) 102983.

[7] A. A. Sodemann, M. P. Ross, B. J. Borghetti, A review of anomaly detection in automated surveillance, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (6) (2012) 1257–1272.

[8] V. Zavrtanik, M. Kristan, D. Skočaj, Reconstruction by inpainting for visual anomaly detection, Pattern Recognition 112 (2021) 107706.

[9] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, A. v. d. 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, 2019, pp. 1705–1714.

[10] H. Park, J. Noh, B. Ham, Learning memory-guided normality for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14372–14381.

[11] Z. Liu, Y. Nie, C. Long, Q. Zhang, G. Li, A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13588–13597.

[12] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, S.-I. Lee, Generative cooperative learning for unsupervised video anomaly detec-

tion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14744–14754.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[14] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[15] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.

[16] J. Zhang, L. Qing, J. Miao, Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 4030–4034.

[17] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, G. Li, Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1237–1246.

[18] B. Wan, Y. Fang, X. Xia, J. Mei, Weakly supervised video anomaly detection via center-guided discriminative learning, in: 2020 IEEE In-

ternational Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.

[19] D. Purwanto, Y.-T. Chen, W.-H. Fang, Dance with self-attention: A new look of conditional random fields on anomaly detection in videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 173–183.

[20] C. Lea, R. Vidal, A. Reiter, G. D. Hager, Temporal convolutional networks: A unified approach to action segmentation, in: European conference on computer vision, Springer, 2016, pp. 47–54.

[21] Y. Cheng, B. Wang, B. Yang, R. T. Tan, Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1157–1165.

[22] C. Cheng, C. Zhang, Y. Wei, Y.-G. Jiang, Sparse temporal causal convolution for efficient action modeling, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 592–600.

[23] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, V. H. C. de Albuquerque, Activity recognition using temporal optical flow convolutional features and multilayer lstm, IEEE Transactions on Industrial Electronics 66 (12) (2018) 9692–9702.

[24] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems (2021).

[25] I. C. Duta, L. Liu, F. Zhu, L. Shao, Improved residual networks for image and video recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 9415–9422.

[26] N. R. Chilkuri, C. Eliasmith, Parallelizing legendre memory unit training, in: International Conference on Machine Learning, PMLR, 2021, pp. 1898–1907.

[27] D. Wang, C. Gong, Q. Liu, Improving neural language modeling via adversarial training, in: International Conference on Machine Learning, PMLR, 2019, pp. 6555–6565.

[28] P. Ma, Y. Wang, J. Shen, S. Petridis, M. Pantic, Lip-reading with densely connected temporal convolutional networks, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2857–2866.

[29] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 233–240.

[30] W. Hao, R. Zhang, S. Li, J. Li, F. Li, S. Zhao, W. Zhang, Anomaly event detection in security surveillance using two-stream based model, Security and Communication Networks 2020 (2020).

[31] M. Z. Zaheer, A. Mahmood, H. Shin, S.-I. Lee, A self-reasoning framework for anomaly detection using video-level labels, IEEE Signal Processing Letters 27 (2020) 1705–1709.

[32] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: European conference on computer vision, Springer, 2020, pp. 322–339.

[33] M. Z. Zaheer, A. Mahmood, M. Astrid, S.-I. Lee, Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection, in: European Conference on Computer Vision, Springer, 2020, pp. 358–376.

[34] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, G. Carneiro, Weakly-supervised video anomaly detection with robust temporal feature magnitude learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4975–4986.

[35] H. Sapkota, Q. Yu, Bayesian nonparametric submodular video partition for robust anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3212–3221.

[36] J.-C. Feng, F.-T. Hong, W.-S. Zheng, Mist: Multiple instance self-training framework for video anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14009–14018.