

**Video Anomaly Detection with U-Net Temporal Modelling and Contrastive
Regularization**

BY

Gan Kian Yu

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2023

REPORT STATUS DECLARATION FORM

Title: Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization

Academic Session: January 2023

I GAN KIAN YU
(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

Address:

No 62, Jalan 3, Taman Telaga Tujuh,

45000 Kuala Selangor,

Selangor

Tan Hung Khoon

Supervisor's name

Date: 24/4/2023

Date: 26/04/2023

Universiti Tunku Abdul Rahman			
Form Title : Sample of Submission Sheet for FYP			
Form Number: FM-IAD-004	Rev No.: 0	Effective Date: 21 JUNE 2011	Page No.: 1 of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TUNKU ABDUL RAHMAN

Date: 24/4/2023

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that **GAN KIAN YU** (ID No: **19ACB01693**) has completed this final year project entitled “Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization” under the supervision of Ts Dr. Tan Hung Khoon (Supervisor) from the Department of Computer Science, Faculty of Information and Communication Technology.

I understand that University will upload softcopy of my final year project in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.


Yours truly,



GAN KIAN YU

DECLARATION OF ORIGINALITY

I declare that this report entitled “**Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :  _____

Name : Gan Kian Yu

Date : 23/4/2023

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisor, Ts Dr Tan Hung Khoon who has given me this bright opportunity to engage in deep learning research project. Thank you for always being patient for guiding me in project implementation and report writing. A million thanks to you.

ABSTRACT

Video anomaly detection (VAD) which is able to automatically identify the location of the anomaly event that happened in the video is one of the current hot study areas in deep learning. Due to expensive frame-level annotation in video samples, most of the VAD are trained with the weakly-supervised method. In a weakly-supervised manner, the labels are at video level. VAD is still an open question and challenging task because the model is trained with a limited sample in weakly supervised video-level labels. In this project, we aim to improve the VAD network with 2 different aspects. Firstly, we explore a technique to model the local and global temporal dependencies. Temporal dependencies are critical to detect anomaly events. Previous methods such as stacked RNN, temporal consistency and ConvLSTM can only capture short-range dependencies. GCN-based methods can model long-range dependencies, but they are slower and more difficult to train. RTFM captures both the short and long-temporal dependencies using two parallel structures, one for each type. However, the two dependencies are considered separately, neglecting the close relationship between them. In this aspect, we propose to use U-Net like structure to model both local and global dependencies for specialized features generation. Second, we explore a new regularization technique in a weakly-supervised manner to reduce overfitting. Insufficient training samples will lead to overfitting easily. Generally, the overfitting issue can be improved by reducing the complexity of the network, data augmentation, injecting noise into the network or applying dropout regularization. For VAD, previous works have applied special heuristics such as sparsity constraint and temporal smoothness to regulate the output of the model. However, none of the existing work has extended a feature-based approach to regularization where the strategy is to learn more discriminative features. In this project, we extend contrastive regularization in a weakly-supervised manner as a new regularization technique to reduce overfitting by learning more discriminative features and enhancing the separability of the features from different classes. We evaluated our model's performance and compared the AUC performance with other state-of-the-art methods. Experimental results show that our model achieves the second-highest AUC performance compared to all published work on a benchmark dataset, namely UCF-Crime using the same pre-trained features.

TABLE OF CONTENTS

TITLE PAGE	i
DECLARATION OF ORIGINALITY	iv
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS	vii-viii
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF SYMBOLS	xi
LIST OF ABBREVIATIONS	xii
CHAPTER 1 INTRODUCTION	1-3
CHAPTER 2 LITERATURE REVIEW	4-13
2.1 Previous Works on Anomaly Detection Algorithm	4-13
2.1.1 Weakly Supervised Anomaly Detection with Multiple Instance Learning (MIL) Frameworks	4-6
2.1.2 Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning	6-8
2.1.3 Localizing Anomalies From Weakly-Labeled Videos	8-10
2.1.4 MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection	10-13
2.2 Limitations of Previous Studies	13
CHAPTER 3 PROPOSED METHOD	14-21
3.1 Modeling Local and Global Temporal Dependencies with U-Net	15-17
3.2 Segment-level Anomaly Classification	17
3.3 Weakly Supervised Contrastive Regularization	18-20
3.4 Loss Function	20-21

CHAPTER 4 Result	21-29
4.1 Experiments and Evaluation	21
4.2 Implementation Details	21-22
4.3 Result on UCF-Crime	22-23
4.4 Comparative and Ablation Study	23-24
4.5 Fine-tuning the model	24-26
4.5.1 Number of Channels	24
4.5.2 Filter Size	24-25
4.5.3 Number of Centers	25
4.5.4 Distance metrics	25-26
4.6 Qualitative Analysis	26-27
4.7 Impact of Contrastive Regularization	28
4.8 Comparison of AUC performance with different category events on UCF-Crime	29
CHAPTER 5 CONCLUSION	30
REFERENCES	31-34
WEEKLY LOG	35-38
POSTER	39
PLAGIARISM CHECK RESULT	40
FYP2 CHECKLIST	42

LIST OF FIGURES

Figure Number	Title	Page
Figure 1.1	General VAD network architecture proposed in previous works	2
Figure 2.1	Network architecture proposed in [1]	5
Figure 2.2	Visualization of computing difference in the mean of top-3 feature magnitudes [2]	7
Figure 2.3	Network architecture of RTFM	7
Figure 2.4	Feature extraction in WSAL	8
Figure 2.5	workflow to compute margin score in WSAL	9
Figure 2.6	workflow to compute margin loss in WSAL	10
Figure 2.7	Network architecture of multiple instance pseudo label generator	11
Figure 2.8	Network architecture of SGA module	12
Figure 3.1	Proposed video anomaly detection with U-Net and Contrastive Regularization	15
Figure 3.2	Modeling local and global dependencies with U-Net	17
Figure 3.3	Supervised contrastive regularization [22] (left), aims to make samples close to the nearest centers of the same class with intra-class compactness and ensure all centers are well separated with inter-class separability. Proposed weakly supervised contrastive regularization (right) to handle weak video-level labels with pseudo-label.	20
Figure 4.1	Predicted anomaly scores for several test videos from UCF-Crime by our proposed method.	26
Figure 4.2	Chronology of events and the corresponding predicted anomaly scores for 2 UCF-Crime test videos.	27
Figure 4.3	Relative distance between segments and the normal versus abnormal centers.	28

LIST OF TABLES

Table Number	Title	Page
Table 4.1	Comparison of AUC performance with other state-of-the-art methods on UCF-Crime.	23
Table 4.2	Ablation studies of our proposed method	23
Table 4.3	Impact of convolutional channel size	24
Table 4.4	Impact of filter size in the residual block	25
Table 4.5	Impact of Number of Centers on UCF-Crime	25
Table 4.6	Impact of Distance metrics in Contrastive Regularization	26
Table 4.7	Comparison of AUC performance with different category events on UCF-Crime.	28

LIST OF SYMBOLS

α	alpha
β	beta
λ	lambda
γ	gamma

LIST OF ABBREVIATIONS

<i>VAD</i>	Video Anomaly Detection
<i>WS-VAD</i>	Weakly Supervised Video Anomaly Detection
<i>MIL</i>	Multiple-Instance Learning
<i>CNN</i>	Convolutional Neural Networks
<i>ConvNet</i>	Convolutional Neural Networks
<i>C3D</i>	Convolutional 3D
<i>I3D</i>	Two-Stream Inflated 3D ConvNet
<i>MIST</i>	Multiple Instance Self-Training Framework
<i>SGA</i>	Self-Guided Attention Module
<i>BN-Inception</i>	Inception with Batch Normalization
<i>RNN</i>	Recurrent Neural Network
<i>LSTM</i>	Long Short-term Memory RNN
<i>ConvLSTM</i>	Convolutional LSTM
<i>ReLU</i>	Rectified Linear Unit

Chapter 1

Introduction

Surveillance camera has been deployed and implemented everywhere but expensive human manual work on live monitoring is required for large-scale deployment surveillance. Therefore, video anomaly detection (VAD) that is able to localize an abnormal event in the surveillance video is one of the hot research topics in deep learning.

Given a video that contains an anomaly event in somewhere the sub-region of the video, the VAD is expected not only to detect the presence of the anomaly event in the video but also to localize the anomaly event that happened in the video. The kind of anomaly event that is expected to be detected by VAD included real-world anomalies such as explosion, robbery, abuse, fighting, shoplifting, stealing, and vandalism. VAD is still an open question and challenging task because the anomaly in the real world come in many different and complex forms.

Due to the infrequency of abnormal events and the difficulty in obtaining enough positive samples, video anomaly detectors are typically trained using unsupervised learning [26,27] or weakly supervised methods [1,2,3,4]. In weakly supervised learning, labels are only provided at the video level, indicating whether a video contains anomalous events that could occur at any point within the video.

The general VAD network architecture is shown in Figure 1.1. It includes 3 major processes which are generic feature extraction with backbone network, anomaly feature extraction and anomaly classification. The purpose of generic feature extraction is to extract all the semantics spatiotemporal features from the video clip. While anomaly feature extraction will extract the specialized feature like anomaly feature from the semantics spatiotemporal features. Then, anomaly classification will predict the anomaly score at the segment level based on the anomaly feature. Based on the anomaly scores, MIL loss will be computed to train the model. In this project, we aim to improve the network with 2 different aspects, including temporal modelling in anomaly feature extraction and regularization to prevent overfitting.

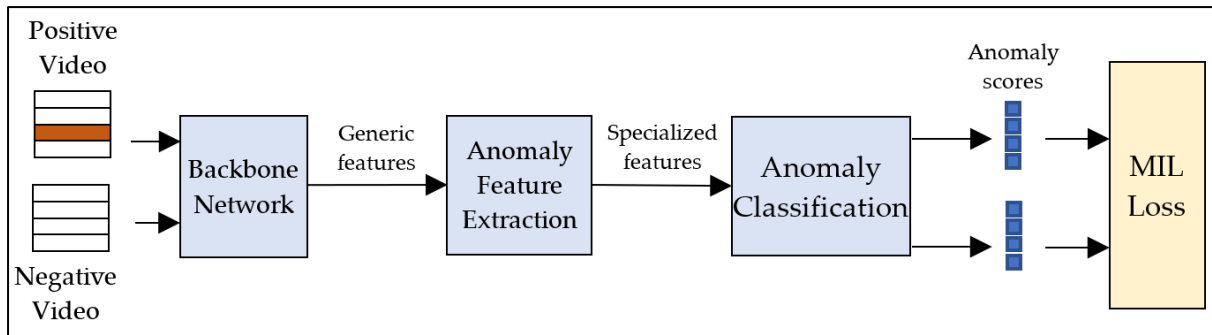


Figure 1.1 General VAD network architecture proposed in previous works

In recent research, it has been proven that modelling local and global temporal dependencies are effective in VAD [2, 29, 30, 31, 32]. Local temporal information captures immediate anomalous characteristics such as sudden movement, unusual human behaviour, and environmental changes while global temporal information provides the overall context that enables the network to distinguish between normal and abnormal scenes in the videos. Previous methods such as stacked RNN [5], temporal consistency [33] and ConvLSTM [34] can only capture short-range dependencies. GCN-based methods [31, 32] can model long-range dependencies, but they are slower and more difficult to train. Although [2] has proposed a method that models local and global temporal dependencies, the structures are considered separately and neglect the close relationship between them. In this paper, we propose a unique approach toward this weakness by using a single structure to model local and global temporal dependencies.

Besides that, we explore a new strategy to reduce overfitting. Overfitting is a common problem due to the scarcity of positive samples available for training. Typically, regularization is achieved by reducing the complexity of the network, adding noise to the network or data, and augmenting the training set [3]. Previous research in VAD has utilized heuristics such as sparsity constraints [1] and temporal smoothness [1] to regulate the network's output. In this work, we adopt a feature-based approach called contrastive regularization where the strategy is to learn more generalizable features. Contrastive regularization [22] has been proven that it can be extended to learn discriminative features and prevent overfitting of the model by enhancing the features of different classes. However, the original work of contrastive regularization is in a supervised manner and none of the previous works has considered using contrastive regularization in a weakly-supervised manner. In this project, we extend the contrastive regularization technique in a weakly-supervised manner.

The main objective of this project is to design a novel anomaly detection system that can model useful temporal dependencies on the relevant part of the video while preventing overfitting. The proposed structure has the following characteristics:

- U-Net structure to model local and global temporal dependencies.
- Contrastive regularization to prevent overfitting.

This project is a video anomaly detection algorithm. The video anomaly detection algorithm is built by neural networks. The video training dataset and testing dataset used in this project is the video anomaly detection benchmark dataset called UCF-Crime. The anomaly event included in the dataset included 13 real-world anomalies such as explosion, arrest, abuse, fighting, shoplifting, stealing, and vandalism. The network will take video data as input for the network and predict the anomaly score for each segment of the video.

There are 2 contributions in this project. Firstly, we extend a U-Net structure with the MIL framework to model the local and global temporal dependencies in a single integrated structure. Secondly, we extend the contrastive regularization technique with MIL framework in a weakly-supervised manner. The experimental result shows that our work achieves state-of-the-art performance on the UCF-Crime benchmark.

Chapter 2

Literature Review

In this chapter, we discuss 4 previous works on anomaly detection algorithms.

2.1 Previous Works on Anomaly Detection Algorithm

In this sub-chapter, we discuss the weakly supervised anomaly detection with Multiple Instance Learning (MIL) framework, Robust Temporal Feature Magnitude Learning (RTFM), Weakly Supervised Anomaly Localization (WSAL), and Multiple Instance Self-Training Framework (MIST).

2.1.1 Weakly Supervised Anomaly Detection with Multiple Instance Learning (MIL) Frameworks

Initially, most previous works proposed a supervised learning method to learn the anomaly. However, preparing a huge number of labelled sample videos is expensive. Also, since the anomaly in the real world come in many different and complex forms, it is challenging to make a complete list of all potential anomalous events. Therefore, in order to avoid spending a lot of time annotating the anomaly and normal label for segments or clips in videos, [1] proposed multiple instances ranking framework to learn anomalies with weakly labelled videos which have only the video-level label. Simply said, the label only shows whether the video contains any anomalous event somewhere in the video but did not show when the anomalous happened in the video. With the objective of predicting a high anomaly score for the positive segment, they applied multiple instance learning (MIL) to create a weakly-supervised learning strategy. As shown in Figure 2.1, they divide the video into segments as an instance of the bag. Then, they predict the anomaly score for each segment. To train the network with weakly labelled video data, they applied a loss function that is able to guide the model to predict a maximum segment score in the anomalous video that is higher than the maximum segment score in the normal video.

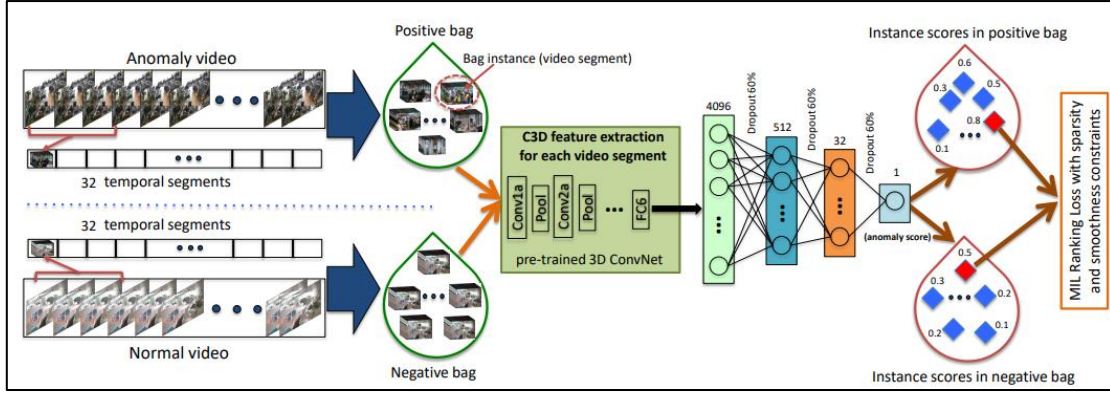


Figure 2.1 Network architecture proposed in [1]

Since their goal is to have the abnormal video segments have greater anomaly scores than the regular video portions, the ranking loss is simply a loss function that rewards abnormal video segments with high scores compared to normal segments. The equation of the loss function is shown in Eq 2.1, where $\max_{i \in \mathcal{B}_a} f(V_a^i)$ and $\max_{i \in \mathcal{B}_n} f(V_n^i)$ represent the maximum segment score in the anomalous video and the maximum segment score in the normal video respectively.

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(V_a^i) + \max_{i \in \mathcal{B}_n} f(V_n^i)). \quad (2.1)$$

In fact, anomaly case in the real world usually occurs for a short amount of time. Therefore, they applied a sparse constraint in the loss function that was also used in [18] and [19] so that the instances (segments) in the anomalous bag should have scores that are sparse, indicating that only a small number of segments may contain the anomaly. In addition, as the video consists of a series of segments, they applied a smoothness constraint in the loss function to minimize the difference in the score between adjacent video segments so that the anomaly score varies naturally between video segments. As a result, they proposed a loss function with sparsity and smoothness constraints as shown in Eq. 2.2, where ① represents the smoothness term and ② indicates the sparsity term.

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(V_a^i) + \max_{i \in \mathcal{B}_n} f(V_n^i)) + \lambda_1 \underbrace{\sum_i^{(n-1)} (f(V_a^i) - f(V_a^{i+1}))^2}_{\text{①}} + \lambda_2 \underbrace{\sum_i^n f(V_a^i)}_{\text{②}}, \quad (2.2)$$

In the training phase, each video frame in anomaly video and normal video will be resized and fixed to a specific frame rate. Then, every video clip will be passed to C3D [6] to extract the features followed by l_2 normalization. The video clips are then divided into segments as an instance of the bag. Each video segment feature will then be obtained by averaging all the clip features in the segment. These features will then pass to a fully connected neural network to

get an anomaly score for each segment. Then, the loss will be computed by using the loss function with sparsity and smoothness constraints in Eq. 2.2.

Although the proposed method successfully timely detects the anomaly frames in anomalous videos with high anomaly scores. However, it failed to identify the normal group activity and generated a false alarm. In other words, the model cannot differentiate the anomalies frames and normal frames well. This issue is mainly caused by 4 major problems. Firstly, the snippet with the highest anomaly score in a positive (anomalous) video may not be one of the positive snippets. Secondly, the training is convergence since the negative (normal) snippet that is used to train the model is randomly selected from the negative video. Thirdly, the training process is not effective since only the top score snippet will be involved in calculating the cost for the model although the video has more than one positive snippet. Lastly, the separation between positive and negative snippets may not be achieved by using a classification score. In addition, the feature extracted is only the feature within 1 segment itself which does not have the temporal feature relation between the neighbours segments. Also, this method does not model any temporal information.

2.1.2 Robust Temporal Feature Magnitude Learning (RTFM)

[2] proposed Robust Temporal Feature Magnitude (RTFM) learning that used a top-k instance instead of a top anomaly score instance to train a model to solve the training issues in [1]. The variable k indicates the number of snippets used to train the model. Instead of predicting anomaly scores based on features extracted by the backbone network proposed in previous work, they proposed a snippet classifier that predicts anomaly scores based on the feature magnitude. Besides 2 loss functions for temporal smoothness and sparsity that are used in previous work, RTFM included 2 additional loss functions called Feature Magnitude Learning and RTFM-enabled Snippet Classifier Learning to train the model.

Feature Magnitude Learning is used to guide the model to update the parameters so that the RTFM model can generate high feature magnitude for positive (anomaly) snippets and low feature magnitude for negative (normal) snippets. The Feature Magnitude Learning loss function is based on the difference in the mean of the top-k feature magnitudes between the positive snippet and negative snippet. An example of the difference in the mean of the top-3 feature magnitudes is shown in Figure 2.2. In Figure 2.2, X^+ and X^- indicate a positive (anomaly) video and a negative (normal) video. $\|x^+\|$ and $\|x^-\|$ indicate feature magnitude for positive (anomaly) snippet and negative (normal) snippet. $\text{Score}(X^+)$ and $\text{score}(X^-)$ indicate the mean of top-k feature magnitude in a positive (anomaly) video and negative (normal) video.

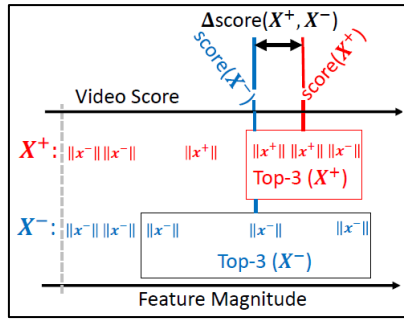


Figure 2.2 Visualization of computing difference in the mean of top-3 feature magnitudes [2]

RTFM-enabled Snippet Classifier Learning is a binary cross-entropy loss function which takes the predicted label for top-k snippets and ground truth video-level label to calculate the loss for the model. The binary cross-entropy loss will guide the model to predict a high anomaly score for a positive snippet and a low anomaly score for a negative snippet.

As shown in Figure 2.3, during the training phase in RTFM, the features extracted by pre-trained networks such as C3D [6] or I3D [7] will be further extracted by the Multi-scale Temporal Network (MTN) with a pyramid of dilated convolutions (PDC) [9] and a temporal self-attention module (TSA) [8]. Then, RTFM will apply the L2 norm to compute feature magnitude and use the top-k feature magnitude to classify the top-k snippets. Then the cost of the network will be computed by feature magnitude learning and RTFM-enabled snippet classifier learning with temporal smoothness and sparsity regularization.

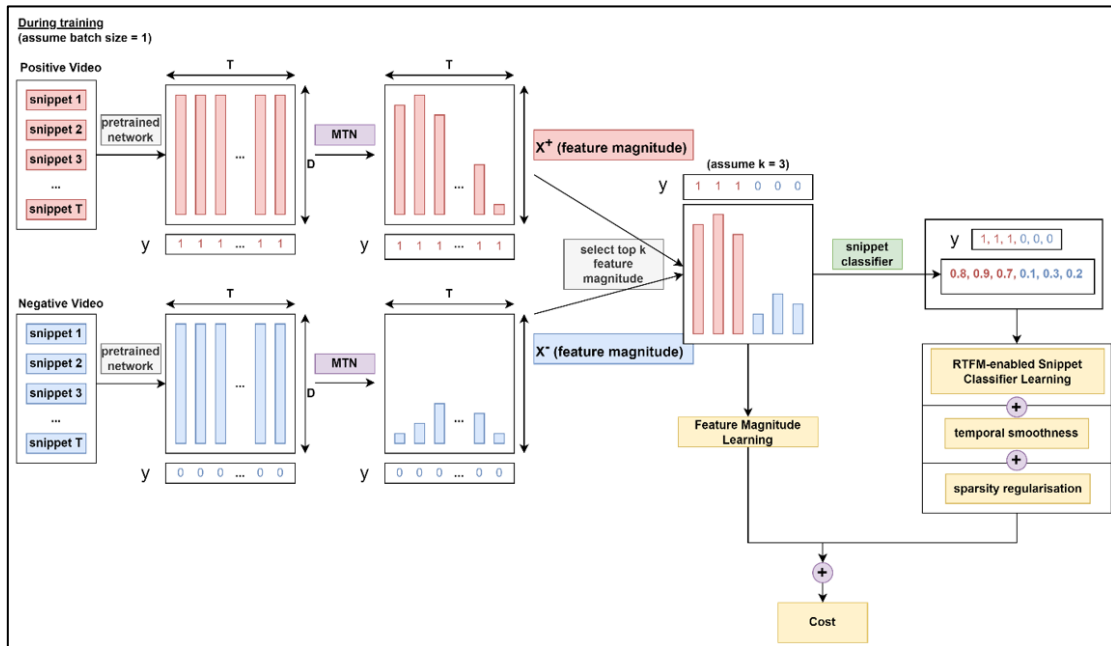


Figure 2.3 Network architecture of RTFM

By using the top-k instance to train the model, there are several weaknesses in MIL that can be solved by RTFM. Firstly, the chance of getting abnormal snippets from positive video (anomaly video) increases since more snippets will be chosen from each video. Secondly, it improves the training convergence since the hard negative (normal) snippets which look like an abnormal snippet for the model will be used to train the model. Thirdly, more abnormal snippets could be included in each abnormal video since more than 1 snippet can be used to train the model. Lastly, a large margin between abnormal and normal snippets can be enforced by using feature magnitude to classify the anomaly since the magnitude can rise throughout the training period.

Although the pyramid of dilated convolutions (PDC) and temporal self-attention module (TSA) are used to capture the local and global temporal dependencies, it is 2 parallel structures that do not have a relationship between the local and global dependencies.

2.1.3 Weakly Supervised Anomaly Localization (WSAL)

Since the previous work has not considered the temporal relation feature among the neighbour segments that bring temporal context for anomaly localization, [3] proposed the Weakly Supervised Anomaly Localization method. As shown in figure 2.4, they used High-order Context Encoding (HCE) model to encode the variation in time series and extract the high-level semantic features based on the feature extracted from the pre-trained model called BN-Inception version of TSN.

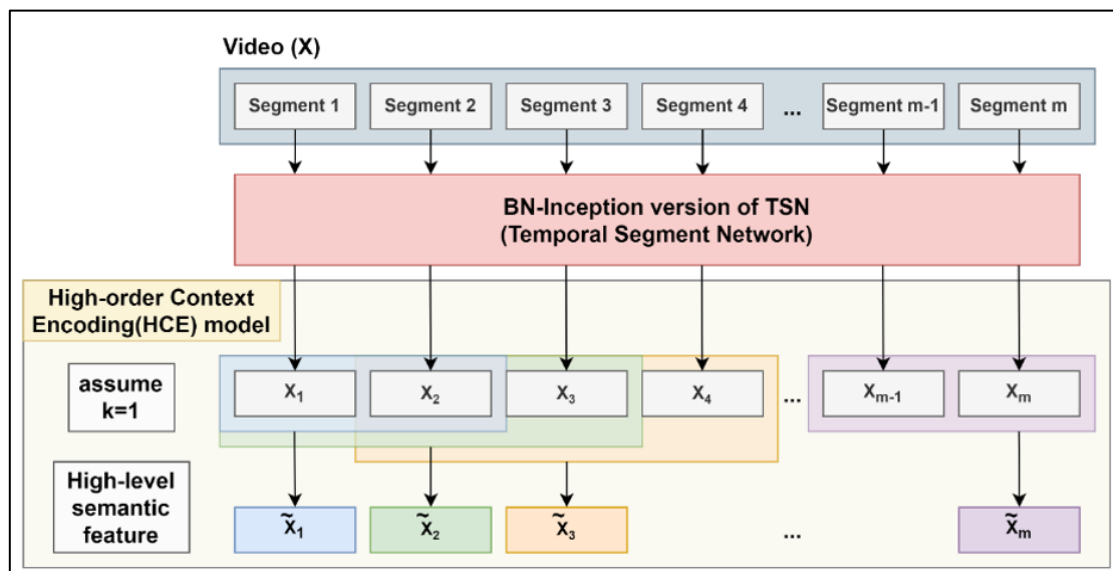


Figure 2.4 feature extraction in WSAL

As shown in figure 2.5, they pass the high-level semantic features to a fully connected layer with a sigmoid activation function to get an immediate semantic score and used cosine

similarity measurement to get a dynamic variation score. The margin dynamic variation score and margin immediate semantic score for the video is then defined by the maximum margin between the 2 scores.

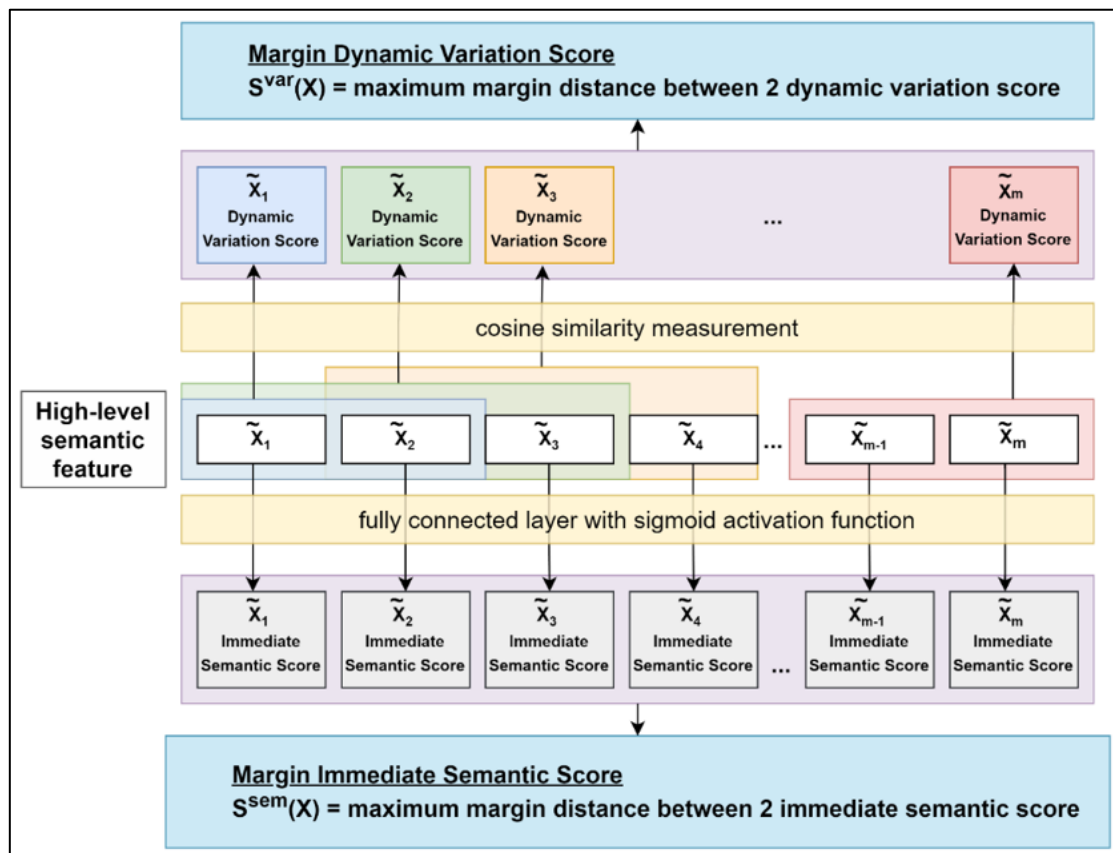


Figure 2.5 workflow to compute margin score in WSAL

Instead of using binary cross entropy to train the classifier that was proposed in previous work, they used a margin distance score to train the classifier. As shown in Figure 2.6, the margin loss for the dynamic variation score or immediate semantic score for a batch is dependent on the mean of the margin score from positive video and negative video respectively. The margin loss will guide the model to predict a large margin score for a positive video and a small margin score for a negative video.

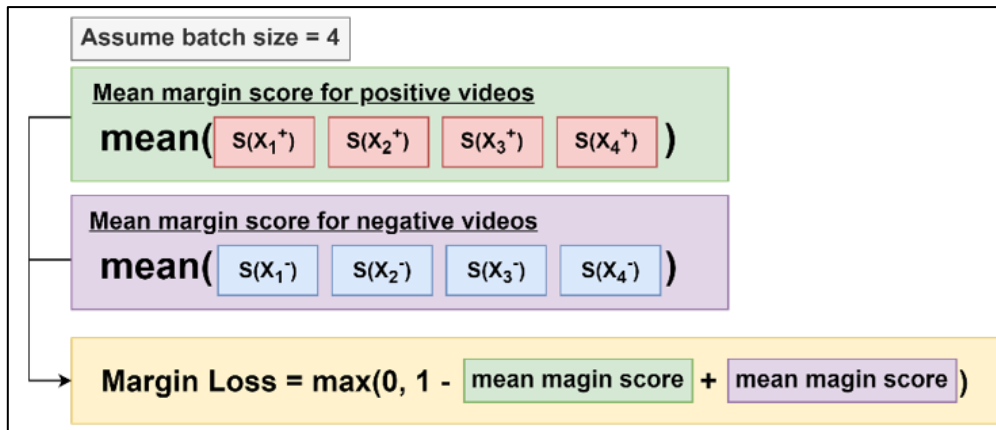


Figure 2.6 workflow to compute margin loss in WSAL

To reduce false alarms happened that due to hardware failure or large changes in the environment, they have also applied noise simulation and hand-crafted anomaly to augment the training video. The augmented videos will be treated as negative training samples and used to train the network. A noise loss function is also applied to guide the network to predict a low immediate semantic score and low dynamic variation score for the sample video that is augmented with noise simulation. Also, a pseudo location loss function is applied to guide the network to predict hand-crafted anomaly samples with lower scores compared to the maximum score from not hand-crafted anomaly samples. In summary, the whole loss function for the network is the aggregation of the margin loss for immediate semantic score and dynamic variation score, sparsity constraint, noise loss and pseudo location loss.

The weakness of this proposed method is that they only have local information but ignored global temporal dependencies.

2.1.4 Multiple Instance Self-Training Framework (MIST)

[4] proposed a multiple instance self-training framework (MIST) that consists of multiple instances of pseudo label generator and self-guided attention-boosted feature encoder. Instead of using the video-level label to train the network, they proposed a Multiple instance pseudo label generator that produces clip-level pseudo label to train the network. While the self-guided attention boosted feature encoder that used a vanilla feature encoder, E a boosted feature encoder, E_{SGA} is used to automatically extract the important region of the feature maps. Also, they used the combination of ranking loss with sparsity constraint and cross-entropy loss function as their proposed framework loss function.

The first step they do in the training phase is the same as previous works, they extract the feature for video clips with C3D [6] / I3D [7] as vanilla feature encoder, E . Then they uniformly sample the video clips to several subsets, L . Instead of fixing the number of segments video

that proposed in previous work, they sample T clips for each subset L . In other words, they treat the temporal length as a hyperparameter to be tuned. After that, they feed the extracted features to a pseudo-label generator to predict the anomaly score for clips, t in each subset, l . Then, they perform average pooling for predicted instance-level anomaly scores in each subset to get the sub-bag score, S_l . The formula to perform average pooling is shown in Eq. 2.3. For all positive videos, they perform temporal smoothing with k neighbours sub-bag for the sub-bag score and followed by min-max normalization to compute pseudo labels in a video. The formula to perform temporal smoothing and min-max normalization are shown in Eq. 2.4. and Eq. 2.5. Then, they combine the pseudo-labelled data with clip-level labelled data to train the proposed feature encoder, E_{SGA} . The network architecture of the proposed multiple instance pseudo label generator is shown in Figure 2.7.

$$S_l = \frac{1}{T} \sum_{t=1}^T s_{l,t}. \quad (2.3)$$

$$\tilde{s}_i^a = \frac{1}{2k} \sum_{j=i-k}^{i+k} s_j^a, \quad (2.4)$$

$$\hat{y}_i^a = \left(\tilde{s}_i^a - \min \tilde{S}^a \right) / (\max \tilde{S}^a - \min \tilde{S}^a), i \in [1, N], \quad (2.5)$$

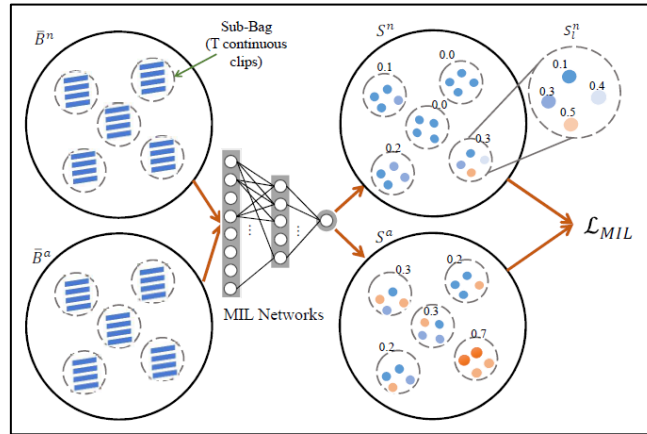


Figure 2.7 Network architecture of multiple instance pseudo label generator

To train the feature encoder, they propose a self-guided attention module (SGA) that used vanilla feature encoder, E as boosted feature encoder, E_{SGA} . Firstly, The 2 feature maps, M_{b-4} and M_{b-5} that generated by 4th and 5th blocks of the vanilla feature encoder, will feed as input to the SGA. Three encoding units, called F_1 , F_2 and F_3 which are created by convolutional layers are included in SGA to encode the feature maps. Secondly, they encode M_{b-4} with F_1 followed by F_2 to generate an attention map, A . Thirdly, they perform element-wise

multiplication for the attention map, A and feature maps M_{b-5} followed by addition with M_{b-5} to get the final attention features map, M_A . The final attention features map, M_A will feed to fully-connected layers, H_c with sigmoid activation to predict the final anomaly score for L_1 . On the other hand, they encode feature maps M_{b-4} with F_1 followed by F_3 into M . Then, they perform channel-wise spatiotemporal average pooling for the normal and abnormal channel, K respectively and followed by the softmax activation function to get a guided anomaly score for the normal class and abnormal class, L_2 . The network architecture of SGA module is shown in figure 2.8.

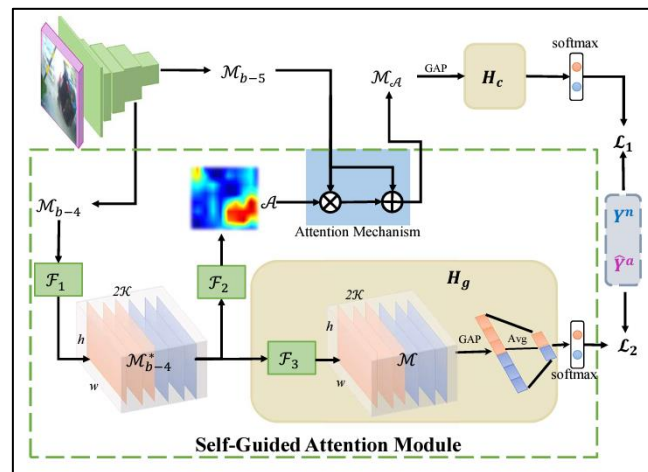


Figure 2.8 Network architecture of SGA module

After the anomaly scores, L_1 and L_2 are obtained, they applied a ranking loss function with sparsity constraint on positive bags and a cross-entropy loss function to train the network. In their ranking loss function, they compared the anomaly score for the highest anomaly score from the positive bag and the highest anomaly score from the negative bag so that the network is always trained by a hard training sample. The ranking loss function is to train the network to predict the highest anomaly score from the positive bag having a greater anomaly score from the negative bag with a margin of ϵ . The ranking loss function is also included a sparsity constraint that included also in [1], [2] and [3]. The ranking loss function with sparsity constraint is shown in Eq. 2.6.

$$\mathcal{L}_{MIL} = \left(\epsilon - \max_{1 \leq i \leq L} S_i^a + \max_{1 \leq i \leq L} S_i^n \right)_+ + \frac{\lambda}{L} \sum_{i=1}^L S_i^a. \quad (2.6)$$

While for the cross entropy loss function, they applied the cross-entropy loss function for both predicted anomaly scores L_1 and L_2 with the pseudo labels for abnormal video and clip-level label for normal video so that the network can predict an anomaly score that is close to 1 for positive sample while predicting anomaly score that is close to 0 for negative sample.

Different from the methods proposed in [1], [2] and [3], they considered the important region feature of the spatial dimension. However, they still ignored the importance of the temporal relation feature.

2.2 Limitations of Previous Studies

Most of the previous studies for surveillance video anomaly detection do not model the local and global temporal dependencies while [2] modelling the local and global temporal dependencies with 2 separate branches. Also, none of the previous works has considered extending the contrastive regularization in the MIL framework to learn discriminative features.

Chapter 3

Proposed Method

Figure 3.1 shows the framework of our proposed network. The objective of the network is to identify abnormal segments in videos with only video-level annotation where segment-level annotations are not available. The training data is a sequence of T segment-level features with dimensions of D_g and $y_i \in \{0, 1\}$ is the corresponding video-level label to indicate whether there are any abnormal segments present in the video. The input features are the generic features extracted from a pre-trained 3D-CNN model called I3D [7].

The segment-level generic features are first input to the U-Net feature extractor to capture the local and global temporal dependencies among the segments in the video. The U-Net features which are enriched with temporal information are then passed to the anomaly classifier to predict segment-level anomalous scores $S_i \in \mathbb{R}^T$ indicating how likely the segment is an abnormal segment. The anomaly classifier also outputs the specialized features $F_i \in \mathbb{R}^{T \times D_f}$ that is used to regularize the network through contrastive regularization.

Since the network is trained with a weakly supervised setup where only video-level annotations are provided, we select k number of segment score $\hat{S} \in \mathbb{R}^k$ and features $\hat{F} \in \mathbb{R}^{k \times D_f}$ based on the segment anomaly score to generate pseudo-labels. For positive videos, the features and anomaly scores of the top- k segments with the highest anomalous scores are selected as pseudo-positive samples and bottom- k segments as pseudo-negative samples. For negative videos, the top- k segments are selected as hard negative normal segments that the network has more difficulty correctly predict them since their anomaly score are high in the situation where they should ideally predict with a low anomaly score. The top- k anomaly scores from positive video and negative video $\hat{S} = \{\hat{S}_i^+, \hat{S}_i^-\}$ are used to compute the binary cross entropy loss to train the model. While the pseudo-label features $\hat{F} = \{\hat{F}_i^+, \hat{F}_i^-\}$ are used to perform contrastive regularization to reduce overfitting.

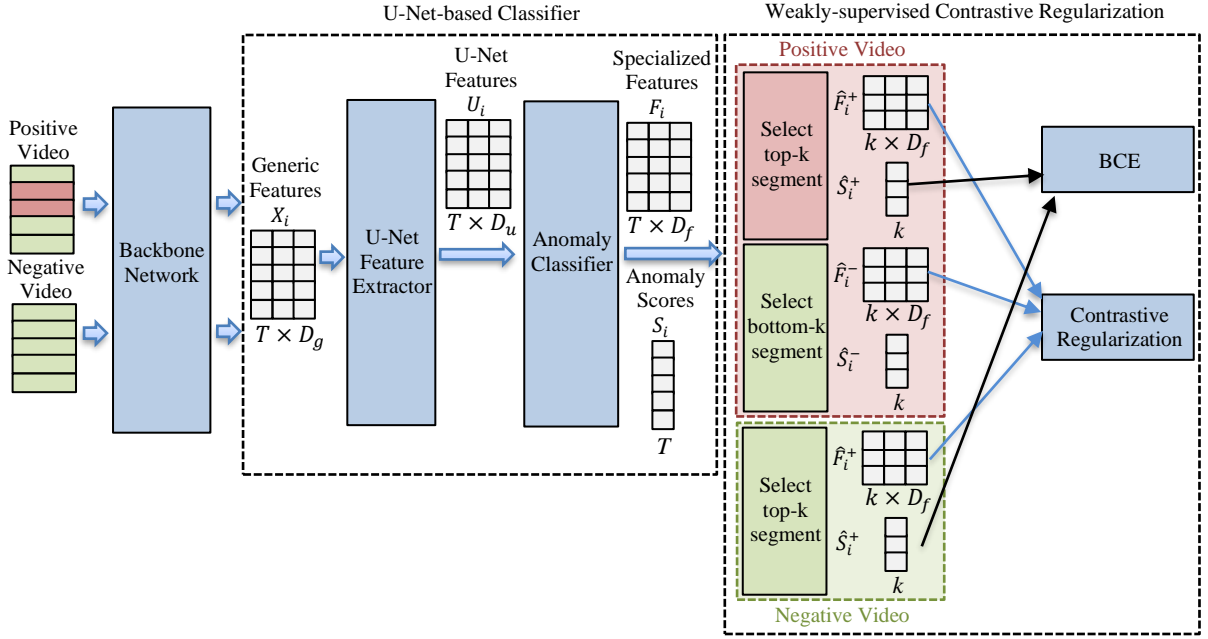


Figure 3.1 Proposed video anomaly detection with U-Net and Contrastive Regularization

3.1 Modeling Local and Global Temporal Dependencies with U-Net

Temporal dependencies have been shown to be critical to the performance of VAD models [2, 29, 30, 31, 32]. Local temporal information captures immediate anomalous characteristics such as sudden movement, unusual human behaviour, and environmental changes while global temporal information provides the overall context that enables the network to distinguish between normal and abnormal scenes in the videos. Although [2] has proposed a method that models local and global temporal dependencies, the structures are considered separately and neglect the close relationship between them. Inspired by this, we proposed U-Net to capture both the local and global temporal dependencies in a single integrated structure.

Although the U-Net is commonly used for image segmentation tasks like medical imaging [20], our research is the first to apply it to detect anomalous segments in videos. Figure 3.2 shows the proposed U-Net adapted for VAD. The network receives a sequence of snippet-level features $X_i \in \mathbb{R}^{T \times D_g}$ from the backbone network as input, and U-Net captures the temporal dependencies among the snippets to enhance the output features $U_i \in \mathbb{R}^{T \times D_u}$. The network has an encoder and a decoder.

The encoder network is responsible for learning both local and global temporal dependencies in the input sequence. Local dependencies are captured using 1-D convolutional operations. Stacking multiple convolutional operations enables the network to learn the global

context in a hierarchical manner. The encoder compresses the input features into a temporally condensed representation and captures local dependencies at the shallower layers and global dependencies at the deeper layers. Due to the fact that the receptive fields in a CNN are more localized at shallower layers and become gradually more global as the network goes deeper [28], the network is able to learn increasingly global information from local ones in previous layers. At the end of the encoding process, the output is a temporally compact representation with high semantic value and global temporal coverage. This encoding process has been demonstrated to be effective in removing noise and capturing common patterns that represent the training samples.

In contrast, the decoder takes the encoded message from the encoder and restores it to segment-level features by combining the activation maps from deeper layers with those of the current layer to obtain a more refined representation, thereby combining both global and local information. To increase the temporal resolution from one layer to the next, transposed convolution is used. The up-sampled features are concatenated with the encoder output at the same height level and then passed to the decoder block for further feature extraction. Through this process, high-level global information is propagated through the layers back to the local segments. As a result, the segment-level features generated by the decoder network are infused with high-level local and global temporal information.

The network height is fixed at $L = 4$, and the temporal resolution $T^{(l)}$ at each height level $l \in \{1, \dots, L\}$ is halved from the previous level, following the formula $T^{(l)} = \frac{T}{2^{l-1}}$. In contrast to conventional U-Net architectures that increase the number of channels with height, our network uses a constant channel size of D_u for all blocks. The network is designed with a residual block structure that consists of three 1-D convolutional layers with ReLU activation and has a skip connection to aid in training. The convolutional layers are set up to produce activation maps of a regular shape of $T^{(l)} \times D_u$ within each block. The conventional U-Net [20] architecture increases the number of channels in the encoder and decreases it in the decoder. However, in our network, the number of channels is kept fixed to D_u in all blocks. This is because our network receives high-level and high-dimensional features from the backbone network, unlike the low-level 3-dimensional features in the conventional U-Net. Also, doubling the channels leads to a huge network and overfitting, while adding a reduction layer results in information loss and lower performance. Therefore, fixing the number of channels to D_u is found to be the best option for enriching the already high-level input features with temporal information.

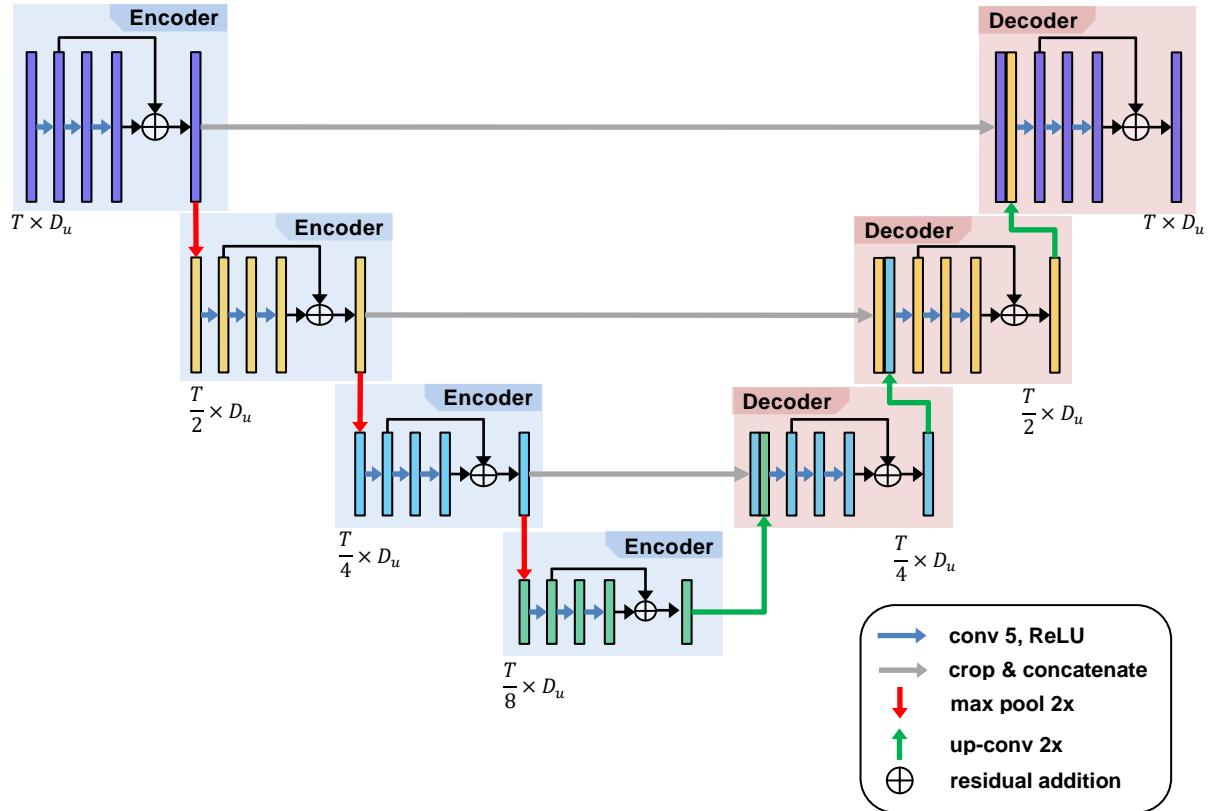


Figure 3.2 Modeling local and global dependencies with U-Net

3.2 Segment-level Anomaly Classification

The anomaly classifier takes the output of the U-Net to predict the anomalous scores $S_i \in \mathbb{R}^T$ for all T segments in the video. An Anomaly classifier is a simple 3-layered multi-layer perceptron (MLP) network. The first and second layers use ReLU activation functions and function as feature extractors, while the last layer uses a sigmoid activation function and functions as a binary classifier to generate anomalous scores for all T segments in the video. The features extracted from the second layer $F_i \in \mathbb{R}^{T \times D_f}$ are subjected to contrastive regularization to reduce overfitting.

3.3 Weakly Supervised Contrastive Regularization

VAD systems face a high risk of overfitting due to the limited number of positive samples. To prevent this, regularization is necessary to enhance the model's ability to generalize. This study proposes a feature-based approach for regularization, where the model is trained to generate features that are robust and can distinguish between events of different types.

In this work, we extend contrastive regularization [22] to a weakly supervised setting. In the weakly supervised setting, we only have access to video-level labels, where a video is labelled as positive if any of its segments is anomalous and negative if none of its segments is anomalous. To deal with the lack of segment-level labels, we use the segment-level anomaly scores generated by the anomaly classification head to generate pseudo-labels using the multi-instance learning (MIL) framework. We extract the top-k segments with the highest anomaly scores as pseudo-positive samples and bottom-k segments as pseudo-negative samples from positive videos. For negative videos, only the top-k segments are selected as hard negative samples. We avoid selecting all segments from negative videos to prevent data imbalance. This approach helps to generate more robust features, making the model more generalizable and noise-tolerant, which is crucial for VAD since positive samples are scarce, making the model vulnerable to overfitting. After generating labels using the anomaly scores, the network can be trained through supervised learning. The multi-instance learning approach assumes that positive segments share common patterns, such as sudden movements, scene changes, or abnormal actions. When selected as pseudo-positive samples, these patterns update the network parameters coherently. Conversely, negative segments in different videos are dissimilar, with varying scenes and activities. If these segments are erroneously selected as pseudo-positive samples, they update the network in an incoherent manner. Over time, the network becomes better at identifying positive segments due to the coherent updates.

Figure 3.3 shows how the contrastive regularization in a supervised setting [22] (left) is extended to our proposed weakly supervised setting (right). The contrastive regularization method enhances the network's generalization by producing distinctive features that distinguish normal features from abnormal ones. This is accomplished by setting up C centers for 2 classes which represent different kinds of normal and anomalous events. These centers are network parameters and will be learned through the training. Let $H = \{H^+, H^-\}$ where $H^+ = \{h_i^+\}_{i=1}^C$ and $H^- = \{h_i^-\}_{i=1}^C$ are the set of all positive and negative centers respectively, the proposed contrastive regularization loss is given by:

$$\begin{aligned}
R_{contrast}(\hat{F}, H) &= \lambda \left(\frac{1}{|\hat{F}^+|} \sum_{f^+ \in \hat{F}^+} \min_{h^+ \in H^+} \|f^+ - h^+\|_2^2 + \frac{1}{|\hat{F}^-|} \sum_{f^- \in \hat{F}^-} \min_{h^- \in H^-} \|f^- - h^-\|_2^2 \right) \\
&+ \beta \frac{1}{C(C-1)} \sum_{h_i \in H} \sum_{h_j \in H, j \neq i} \max(0, m - \|h_i - h_j\|_2^2)
\end{aligned}$$

Where λ is the compactness strength, β is the separability strength, $|\bullet|$ is the cardinality of a set, and $\|\bullet\|_2^2$ is the L2-norm. The first term is the intra-center compactness which minimizes the distance between the feature f_i and its closest center in the same class h_i . Both the pseudo-positive and pseudo-negative segments from positive videos are pulled towards different types of class centers to enable distinguishing between abnormal and normal events within the same video.

The second term is the inter-class separability that enforces the network to separate all the centers. When the distance of any 2 centers is smaller than the margin value m , the inter-class separability will incur a cost. The inter-class separability promotes the learning of a more diverse set of features, where each feature represents a different type of anomaly or normal event. Furthermore, the network's classification decision can be associated with the events associated with the nearest class center, which enhances the explainability of the network. The learned centers play a crucial role in generalizing the different common patterns within each class and distinguishing the two classes. Therefore, when the network generates features that are close to these centers, it helps to reduce overfitting.

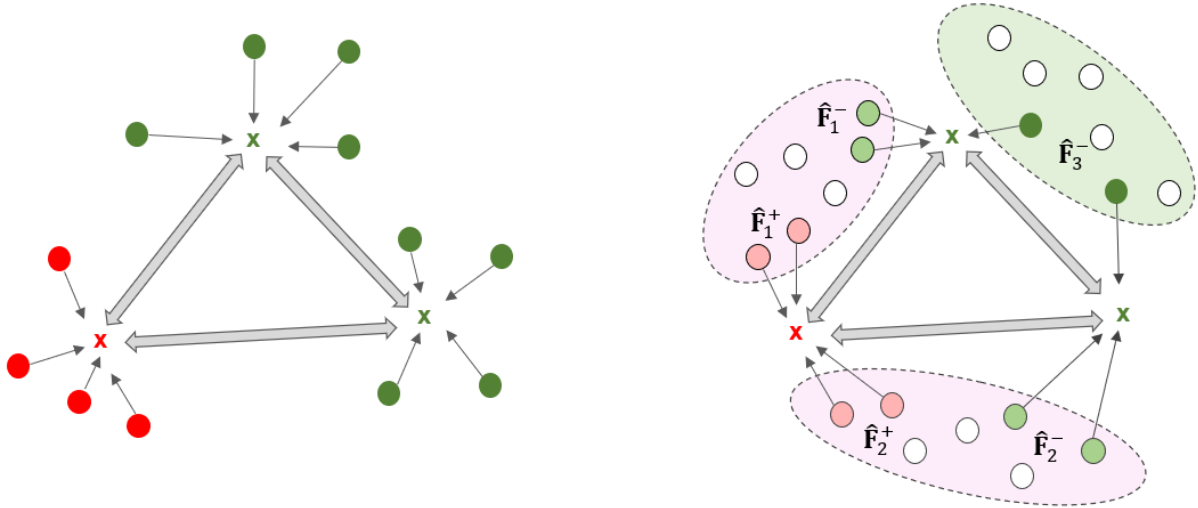
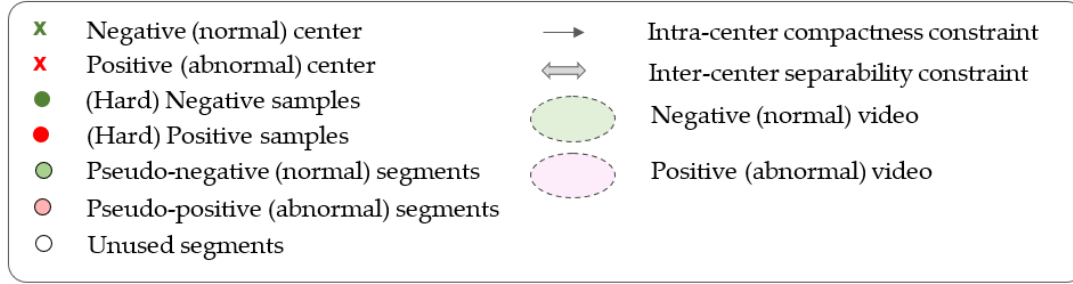


Figure 3.3 Supervised contrastive regularization [22] (left), aims to make samples close to the nearest centers of the same class with intra-class compactness and ensure all centers are well separated with inter-class separability. Proposed weakly supervised contrastive regularization (right) to handle weak video-level labels with pseudo-label.

3.4 Loss Function

With anomaly score for T segments $S = \{S_1, S_2, \dots, S_T\}$, anomaly score for selected top-k segment $\hat{S} = \{\hat{S}_i^+, \hat{S}_i^-\}$, features of the pseudo-label segment $\hat{F} = \{\hat{F}_i^+, \hat{F}_i^-\}$ and centers $H = \{H^+, H^-\}$, the overall loss function of the network is defined as follows:

$$L_{overall} = L_{BCE}(\hat{S}) + R_{contrast}(\hat{F}, H) + \gamma \sum_i^{(T-1)} (S_i - S_{i+1})^2 + \alpha \sum_i^T S_i$$

Where $L_{BCE}(\hat{S})$ is the binary cross entropy loss which minimizes the data loss with selected top-k segment to train the VAD model. $R_{contrast}(\hat{F}, H)$ is the contrastive regularization to regulate the network and prevent overfitting. The third term is the temporal smoothness constraint that is used to force the model to predict a low difference in segment score compared to the neighbours' segment score since the event in the real world is changed slowly. The fourth term is the sparsity constraint that used to force the model to predict only a few segments with

high segment scores since the anomaly event in the real world is often occurs in a short period. The γ and α are the hyperparameters to adjust the strength of temporal smoothness and sparsity constraint.

CHAPTER 4 Result

4.1 Experiments and Evaluation

UCF-Crime [1] is a large-scale dataset that consists of long untrimmed real-world weakly labelled surveillance video data. The duration of the sample included from 8 seconds to 9 hours. The dataset consists of 950 long untrimmed surveillance videos that captured the anomaly event and 950 normal videos. Besides, UCF-Crime have 290 video data with frame-level labels.

Similar to previous studies on VAD [1,2,3,4,26,27], the performance of our proposed system is evaluated using the frame-level AUC metric, which quantifies the area under the ROC curve. A higher AUC indicates better performance. For AUC computation, the anomaly score at the segment level is extended to all frames within the segment. In addition to quantitative evaluation, we also provide qualitative results to assess the localization performance of our system and the impact of contrastive regularization.

4.2 Implementation Details

We applied data augmentation to the dataset. We do 10-crop augmentation for both training data and testing data. The cropped frame size is 210x280 pixels which is 87.5% of the original size. As with other weakly-supervised methods, for every 16-frame video clip, we used pre-trained I3D [7] as a feature extractor to extract 1024D features followed by l2 normalization. For train data, we further equally divide the clips into 32 segments and average all 16-frame clip features within a segment to get the features for a segment. For short videos that have less than 32 clips, we insert some blank frames at the end of the video. In the end, the features of the train data are at the segment level with a uniform temporal resolution while the features of the test data are at the clip level with a varying temporal resolution. Each mini-batch included multiple 32 normal segments and 32 anomaly segments. With training and testing data ready, we build our proposed neural network and train the model.

For the U-Net feature extractor, we set the number of channels D_u to 1024 for all blocks. The number of neurons for each layer in the classification head is 512 units, 32 units and 1 unit. We applied 70% dropout regularization [23] between the fully connected layers. ReLU

Bachelor of Computer Science (Honours)
Faculty of Information and Communication Technology (Kampar Campus), UTAR

[21] activation is also used for the first and second layers while Sigmoid activation is used in the last fully connected layer to predict the score in the range from 0 to 1. With a learning rate of 0.0001, the model is trained using the Adam optimizer [24] with a weight decay of 0.01 and batch size of 64 for 500 epochs. As [1], we set the hyperparameter of temporal smoothness γ and sparsity constraints α to 0.0008. For the hyperparameter in the multicenter loss function, we set the number of centers per class $C=16$, $\lambda = 0.2304$, $\beta = 0.00256$ and $m = 1.25$.

4.3 Results on UCF-Crime

The comparison of AUC performance with other state-of-the-art methods on UCF-Crime [1] is shown in Table 4.1. We compared our proposed network to unsupervised methods [26,27] and weakly supervised methods [1,2,3,4]. Overall, the methods that use weakly supervised methods tend to perform much better than those using unsupervised methods. This suggests that having some form of supervisory labels, even if they are weak, it is crucial for achieving better results.

Our proposed model's AUC performance surpasses most of the SOTA methods with an AUC of 85.24% using the same I3D [7] features. Our proposed method outperforms MIL-ranking [1] by 7.32%, RTFM [2] by 0.94% and MIST [4] by 7.32%. Among all the methods that were evaluated, HCE [3] achieved the highest AUC of 85.38%. However, the performance has come from its data augmentation method that included hand-crafted anomalies (HC) and noise simulation (NS) in their training dataset. Without the data augmentation method, the performance of HCE [3] drops to 84.44%. Therefore, our proposed method actually outperforms HCE [3] by 0.71% on the same training dataset. This is mainly due to the efficacy of U-Net that able to capture both local and global temporal information compared to the temporal modelling method in HCE [3] that mainly captures local temporal information. On the other hand, while RTFM can capture both types of dependencies, they are implemented in two independent structures and cannot effectively model the interaction between the two. In contrast, U-Net can model both types of dependencies more effectively, which results in better performance compared to other methods.

Table 4.1. Comparison of AUC performance with other state-of-the-art methods on UCF-Crime.

Supervision	Method	Feature	AUC (%)
Unsupervised	Sparsity-Combination [27]	C3D RGB	65.51
	BODS [26]	I3D RGB	68.26
	GODS [26]	I3D RGB	70.46
Weakly supervised	MIL-ranking [1]	C3D RGB	75.41
	MIST [4]	C3D RGB	81.40
	RTFM [2]	C3D RGB	83.28
	MIL-ranking [1]	I3D RGB	77.92
	MIST [4]	I3D RGB	82.30
	RTFM [2]	I3D RGB	84.30
	HCE [3] (original training dataset)	I3D RGB	84.44
HCE [3] (with noise-augmented dataset)	I3D RGB	85.38	
	Ours	I3D RGB	85.24

4.4 Comparative and Ablation Study

We perform the ablation study as shown in Table 4.2 to evaluate the effectiveness of our proposed method. The baseline model represents the 3-layered MLP network from [1] that does not have any temporal modelling in the network. The second model is the network with the pyramid of dilated convolution (PDC) and transformer structure (TSA) from [2] that models the local and global temporal dependencies separately in 2 branches.

Table 4.2. Ablation studies of our proposed method

Baseline	PDC+TSA [2]	U-Net	Contrastive Regularization	AUC (%)
✓				81.74
✓	✓			82.20 (+0.46)
✓		✓		83.43 (+1.23)
✓		✓	✓	85.24 (+1.81)

The baseline model achieves the lowest AUC of 81.74% due to a lack of temporal modelling. When the baseline model is extended with PDC and TSA, the AUC improves to 82.20%. This show that the performance of VAD can be improved by learning temporal dependencies. However, the PDC+TSA approach only models local and global temporal dependencies independently, without considering their close relationship. The proposed U-Net feature extractor resolves this weakness by modelling both local and global temporal dependencies in a single structure. By extending the baseline network with the U-Net feature extractor, the AUC performance is boosted to 83.43%.

Next, we evaluate the effectiveness of contrastive regularization for VAD. The AUC performance was boosted to 85.24% when extending both the U-Net feature extractor and contrastive regularization to the baseline network. This shows that networks that learn to separate the feature are indeed more generalizable.

4.5 Fine-tuning the model

4.5.1 Number of Channels

In this section, we evaluate the impact of channel and filter size in the U-Net feature extractor and the number of centers in contrastive regularization. To study the impact of the channel sizes D_u in U-Net, we evaluate the following channel sizes: 256, 512, 1024 and 2048. Different from traditional U-Net, in our design, all convolutional layers have the same channel size. Since the input to U-Net from the backbone network has a channel size of 1024, the first 2 settings ($D_u = 256$ and $D_u = 512$) will shrink the channel size while the last setting ($D_u = 2048$) will expand the channel size.

Table 4.3. Impact of convolutional channel size

Channel Size, D_u	AUC (%)
256	84.08
512	84.85
1024	85.24
2048	83.78

The model with the same channel size as the backbone network feature ($D_u = 1024$) achieved the best AUC. The lower channel size settings ($D_u = 256$ and $D_u = 512$) having the lower AUC performance might be due to the loss of useful information when compressing the channel size. While the higher channel size setting ($D_u = 2048$) also has a lower AUC performance because a bigger network having higher parameters needed to be tuned. Therefore, it is more difficult to optimize and easier to overfit without enough training samples.

4.5.2 Filter Size

Next, we evaluate the impact of filter size in U-Net. The filter sizes of 3, 5 and 7 are evaluated. A suitable filter size is important so that the network can effectively capture the temporal dependencies information. As shown in Table 4.4, the model with filter size $f = 5$ has the best performance. A larger filter size $f = 7$ makes the network less sensitive to subtle local cues. A larger filter size also causes the network to have more parameters and therefore it is easier to overfit when the training samples are not sufficient. A smaller filter size $f = 3$ cannot

effectively capture longer-range temporal dependencies, especially at the current depth level due to its limited temporal range.

Table 4.4. Impact of filter size in the residual block

Filter Size, f	AUC (%)
3	83.44
5	85.24
7	83.84

4.5.3 Number of Centers

In this section, we evaluate the impact of the number of centers in our proposed contrastive regularization method. The number of centers C in contrastive regularization signifies the number of representative events captured by the model. We evaluate $C = 0, 2, 4, 8, 16$ and 24 . The contrastive regularization is disabled when $C = 0$. As shown in Table 4.5, the performance of the model improved when contrastive regularization is enabled ($C \geq 2$). The performance of the model improved significantly by 2.1% to 85.24% with the number of centers $C = 16$. The performance does not have further improvement by further increasing the number of centers to 24 because the network is more difficult to train, and 16 centers are sufficient to explain the anomalies.

Table 4.5. Impact of Number of Centers on UCF-Crime

Number of Centers, C	AUC (%)
0	83.14
2	84.01
4	84.09
8	84.01
16	85.24
24	84.89

4.5.4 Distance metrics

In explore the best formula for contrastive regularization, we experiment and evaluate the contrastive regularization using different distance metrics. Specifically, the Euclidean distance and cosine similarity are used to measure the difference between feature-center and center-center respectively. As shown in Table 4.6, the contrastive regularization with Euclidean distance outperforms the one with cosine similarity by 1.66%. This is because Euclidean distance measure also the magnitude of the vectors compared to cosine similarity only measures the angle between the vectors.

Table 4.6. Impact of Distance metrics in Contrastive Regularization

Distance metrics	AUC (%)
Euclidean distance	85.24
Cosine similarity	83.58

4.6 Qualitative Analysis

In this section, we perform a qualitative analysis of the model. Figure 4.1 shows the graph of anomaly score versus frame number for several test videos by our model. The red colour region indicates the region where an anomaly event happened. Figure 4.1 (a) – (e) shows the result for 6 positive videos that contain anomalous segments such as arson, road accident, robbery, shooting and shoplifting. Figure 4.1 (f) shows a normal video without any anomaly event while (g) and (h) show 2 failure cases where (g) is a missed anomaly detection case and (h) is a false alarm case. In general, the anomaly scores produced by the network accurately correspond to the ground truth where high anomaly scores are generated for abnormal regions and low scores for normal regions. Figure 4.1 (f) shows that the network is able to correctly produce low anomaly scores for every frame of a normal video, indicating that it is accurately detecting normal regions. Figure 4.1 (g) shows that the network failed to detect a shop-lifting event, where a man was caught on camera putting a stolen watch into his pocket. The reason for the failure was due to the subtle nature of the action, as well as the fact that the stolen item was occluded, making it a challenging detection even for a human observer who is not paying close attention. Figure 4.1 (h) shows a false alarm where the network incorrectly identifies a group of people making contact with each other as a fighting event.

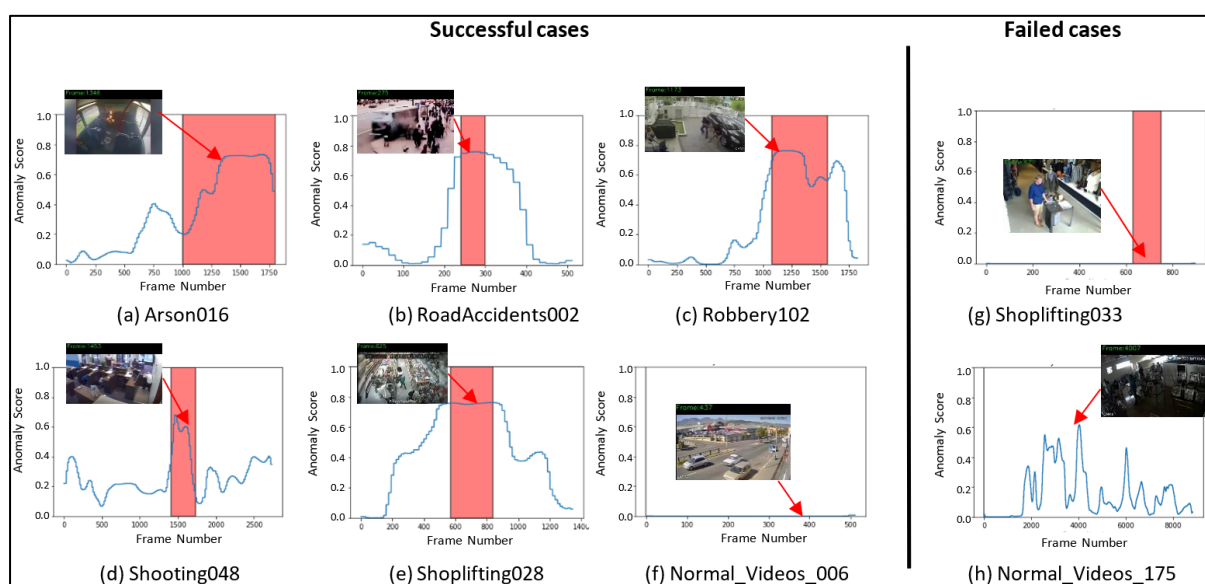


Figure 4.1 Predicted anomaly scores for several test videos from UCF-Crime by our proposed method.

Figure 4.2 provides a more detailed analysis of two test videos that involve burglary and explosion events. The first video, titled "Burglary037" begins with an empty cashier counter shown in Frame 78. A man then enters and climbs over the counter in Frame 257. Then, he passes several stolen bottles of wine to his accomplice in Frame 796. The man then proceeds to ransack the cash register before leaving quickly in Frame 1815. Our proposed network accurately identifies the initial scene as normal and consistently produces higher anomaly scores after the appearance of the burglary. The model also predicts a lower anomaly score after the burglars left. The second video, titled "Explosion033," included two separate explosion events in a single video. The anomaly scores are lower for the scenes leading up to the two explosions at their respective locations (frames 499 and 1505) but become significantly higher when the explosions occur (frames 1075 and 2095). This demonstrates that the model can detect anomalous events within a frame, even though it was trained using video-level labels.

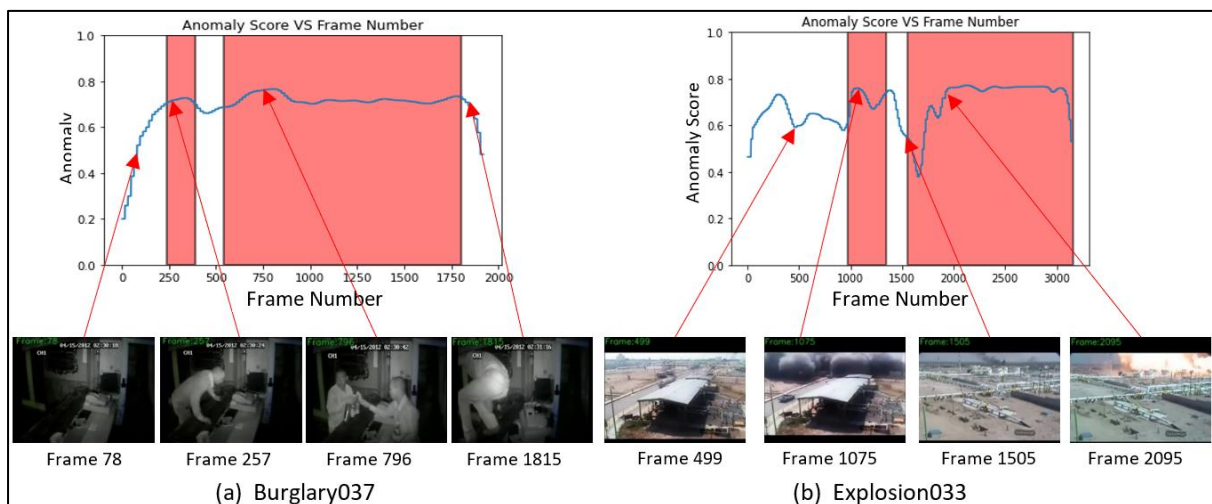


Figure 4.2 Chronology of events and the corresponding predicted anomaly scores for 2 UCF-Crime test videos.

4.7 Impact of Contrastive Regularization

To evaluate the effectiveness of contrastive regularization, we compare the distance between a segment feature \mathbf{f} to its nearest normal center $\mathbf{h}^- \in \mathbf{H}^-$ and its nearest anomaly center $\mathbf{h}^+ \in \mathbf{H}^+$. The relative distance is computed as follows:

$$Dist(\mathbf{f}, \mathbf{H}) = \min_{\mathbf{h}^- \in \mathbf{H}^-} \|\mathbf{f} - \mathbf{h}^-\|_2^2 - \min_{\mathbf{h}^+ \in \mathbf{H}^+} \|\mathbf{f} - \mathbf{h}^+\|_2^2$$

Figure 4.6 shows the relative distances of the segments in 8 different test videos. A negative value of relative distances indicated that \mathbf{f} is closer to the center of normal activity than to the center of abnormal activity. Conversely, if the value relative distances are positive, it indicates that \mathbf{f} is closer to the center of abnormal activity. The red colour region indicates the region where an anomaly event happened. Figure 4.6(a) – (e) shows 5 different test videos that contain anomalous events while figure 4.6(f) represents a test video that does not have any anomalous events. The embedding feature of the normal event, which is represented by the white region, is situated closer to the normal center (below the horizontal line). In contrast, the embedding feature of the anomalous event (red zone) is located closer to the anomaly center (above the horizontal line). This indicates that the features generated using contrastive regularization are more effective at distinguishing between the two event types and are well-aligned with the correct event category. Figure 4.3(g) – (h) shows 2 failure cases on a test video that contains an anomalous event and a test video that does not, respectively. For the former, the features generated are unable to accurately capture the subtle anomalous event. For the latter, the features generated indicate that normal actions are occasionally misidentified as abnormal due to the high level of activity in the scene.

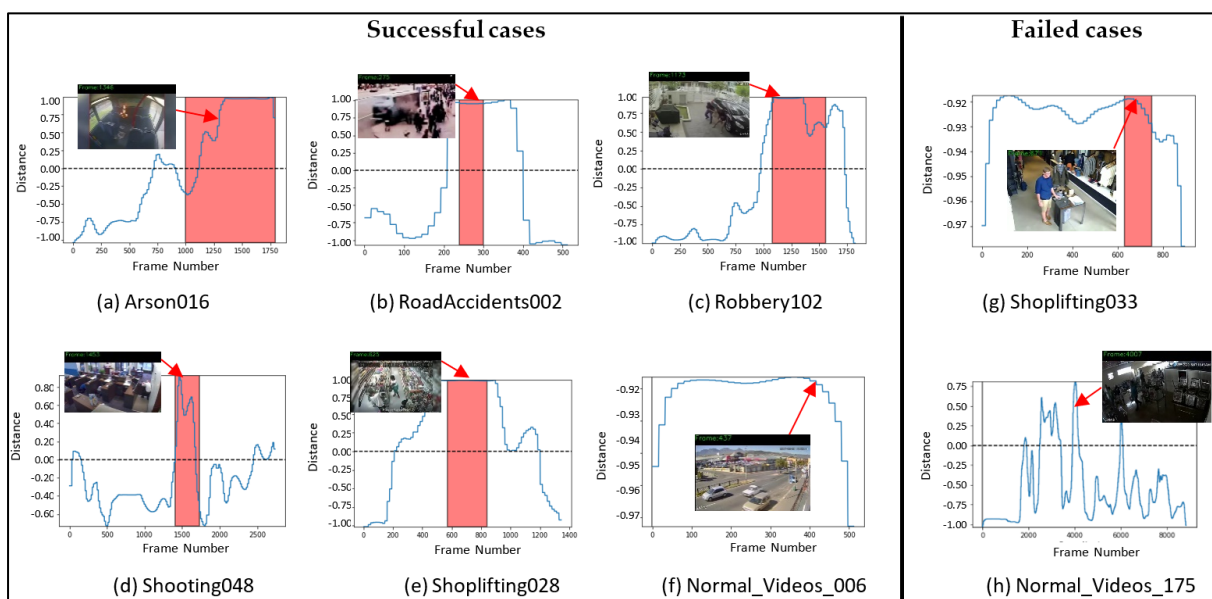


Figure 4.3 Relative distance between segments and the normal versus abnormal centers.

4.8 Comparison of AUC performance with different category events on UCF-Crime

In this section, we evaluate our model's performance for every different category event on UCF-Crime. Since the ground truth of normal video is all 0, it is not applicable to compute the AUC. Table 4.7 shows the AUC performance for different category events sorted by AUC. As shown in Table 4.7, there are a total of 13 categories of anomaly events are evaluated. In general, the lower the ratio of train video and test video ($\frac{\text{number of train video}}{\text{number of test video}}$), the lower the AUC performance due to the insufficient training sample. For anomaly cases that happened without humans appearing on the scene like arson, explosion and road accident, the model is only able to localize the anomaly event based on the understanding of the environment. Therefore, the model has difficulty localizing such anomaly cases because the background and environment in UCF-Crime are very different for each sample. While anomaly cases like abuse, vandalism, assault and fighting, the model is able to understand the event based on the understanding of human action. The more unique and distinct the body movements will have appeared for a particular anomaly case, the easier the model can understand and localize the anomaly event.

Table 4.7. Comparison of AUC performance with different category events on UCF-Crime.

Category	AUC (%)	Number of Test Video	Number of Train Video
Abuse	0.8767	2	48
Vandalism	0.8560	5	45
Assault	0.8311	3	47
Fighting	0.8292	5	45
Stealing	0.8038	5	95
Robbery	0.7323	5	145
Burglary	0.7293	13	87
Shooting	0.7168	23	27
Arrest	0.6837	5	45
Shoplifting	0.6559	21	29
Arson	0.5868	9	41
Explosion	0.5015	21	29
RoadAccidents	0.4865	23	127

CHAPTER 5

Conclusion

In conclusion, this study proposes the use of U-Net to effectively detect anomalous segments in a video by simultaneously capturing both local and global temporal information. Unlike previous methods that separately model these dependencies, U-Net learns global temporal dependencies in the encoder and propagates this information back to the local level in the decoder. To address overfitting, a weakly supervised contrastive regularization approach is introduced, which promotes discriminative and generalizable feature representations. For future work, it is interesting to explore how the encoder and self-attention mechanism in the transformer model can be integrated with our proposed method to further improve the performance of the model.

REFERENCES

- [1] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” *arXiv.org*, 14-Feb-2019. [Online]. Available: <https://arxiv.org/abs/1801.04264>. [Accessed: 03-Sep-2022].
- [2] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, “Weakly-supervised video anomaly detection with robust temporal feature magnitude learning,” *arXiv.org*, 06-Aug-2021. [Online]. Available: <https://arxiv.org/abs/2101.10030>. [Accessed: 03-Sep-2022].
- [3] H. Lv, C. Zhou, C. Xu, Z. Cui, and J. Yang, “Localizing anomalies from weakly-labeled videos,” *arXiv.org*, 14-Apr-2021. [Online]. Available: <https://arxiv.org/abs/2008.08944>. [Accessed: 03-Sep-2022].
- [4] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, “Mist: Multiple instance self-training framework for video anomaly detection,” *arXiv.org*, 04-Apr-2021. [Online]. Available: <https://arxiv.org/abs/2104.01633>. [Accessed: 03-Sep-2022].
- [5] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked RNN framework,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 341–349.
- [6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” *arXiv.org*, 07-Oct-2015. [Online]. Available: <https://arxiv.org/abs/1412.0767>. [Accessed: 03-Sep-2022].
- [7] J. Carreira and A. Zisserman, “Quo Vadis, action recognition? A new model and the kinetics dataset,” *arXiv.org*, 12-Feb-2018. [Online]. Available: <https://arxiv.org/abs/1705.07750>. [Accessed: 03-Sep-2022].
- [8] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local Neural Networks,” *arXiv.org*, 13-Apr-2018. [Online]. Available: <https://arxiv.org/abs/1711.07971>. [Accessed: 03-Sep-2022].

- [9] X. Wang, Y. Zhao, T. Yang, and Q. Ruan, "Multi-scale context aggregation network with attention-guided for crowd counting," *arXiv.org*, 06-Apr-2021. [Online]. Available: <https://arxiv.org/abs/2104.02245>. [Accessed: 03-Sep-2022].
- [10] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *arXiv.org*, 08-May-2017. [Online]. Available: <https://arxiv.org/abs/1705.02953>. [Accessed: 03-Sep-2022].
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv.org*, 10-Dec-2015. [Online]. Available: <https://arxiv.org/abs/1512.03385>. [Accessed: 23-Nov-2022].
- [12] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-kin Wong, and W.-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *arXiv.org*, 19-Sep-2015. [Online]. Available: <https://arxiv.org/abs/1506.04214>. [Accessed: 03-Sep-2022].
- [13] A. Nagpal and G. Gabrani, "Python for Data Analytics, Scientific and Technical Applications," 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 140-145, doi: 10.1109/AICAI.2019.8701341.
- [14] C. D. Costa, "Reasons to choose pytorch for deep learning," *Medium*, 06-Apr-2020. [Online]. Available: <https://towardsdatascience.com/reasons-to-choose-pytorch-for-deep-learning-c087e031eaca>. [Accessed: 04-Sep-2022].
- [15] W. Li and N. Vasconcelos, "Multiple instance learning for soft bags via top instances," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4277-4285, doi: 10.1109/CVPR.2015.7299056.
- [16] C. B. Vennerød, A. Kjærran, and E. S. Bugge, "Long short-term memory RNN," *arXiv.org*, 14-May-2021. [Online]. Available: <https://arxiv.org/abs/2105.06756>. [Accessed: 04-Sep-2022].
- [17] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and long short-term memory (LSTM) network," *arXiv.org*, 31-Jan-2021. [Online]. Available: <https://arxiv.org/abs/1808.03314>. [Accessed: 04-Sep-2022].

- [18] C. Lu, J. Shi and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2720-2727, doi: 10.1109/ICCV.2013.338.
- [19] B. Zhao, L. Fei-Fei and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," *CVPR 2011*, 2011, pp. 3313-3320, doi: 10.1109/CVPR.2011.5995524.
- [20] U-Net Ronneberger, O., Fischer, P., & Brox, T, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv.org*, 18-May-2015. [Online]. Available: <https://doi.org/10.48550/arXiv.1505.04597>. [Accessed: 23-Nov-2022].
- [21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," *OpenReview*, 01-Jan-2010. [Online]. Available: <https://openreview.net/forum?id=rkb15iZdZB>. [Accessed: 04-Sep-2022].
- [22] M. Tanveer, H. -K. Tan, H. -F. Ng, M. K. Leung and J. H. Chuah, "Regularization of Deep Neural Network With Batch Contrastive Loss," in *IEEE Access*, vol. 9, pp. 124409-124418, 2021, doi: 10.1109/ACCESS.2021.3110286.
- [23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., "Dropout: a simple way to prevent neural networks from overfitting", *The journal of machine learning research*, 2014. [Online]. Available: [https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer,.](https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer,) [Accessed: 29-Nov-2022].
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv.org*, 30-Jan-2017. [Online]. Available: <https://arxiv.org/abs/1412.6980>. [Accessed: 29-Nov-2022].
- [25] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

- [26] J. Wang and A. Cherian, “GODS: Generalized one-class discriminative subspaces for anomaly detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8200–8210.
- [27] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 FPS in MATLAB,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 2720–2727.
- [28] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, “Do vision transformers see like convolutional neural networks?” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2021, pp. 12116–12128.
- [29] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked RNN framework,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 341–349.
- [30] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—A new baseline,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6536–6545.
- [31] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Aug. 2020, pp. 322–339.
- [32] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, “Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1237–1246.
- [33] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—A new baseline,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6536–6545.
- [34] W. Liu, W. Luo, Z. Li, P. Zhao, and S. Gao, “Margin learning embedded prediction for video anomaly detection with a few anomalies,” in Proc. 28th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 3023–3030.

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3 Year 3	Study week no.: 2 - 4
Student Name & ID: Gan Kian Yu & 19ACB01693	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- U-Net feature extractor and contrastive regularization
- spatial-attention module

2. WORK TO BE DONE

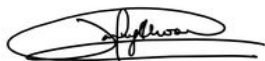
- generate a web-based demo
- analyze results per crime category
- apply contrastive regularization with different distance metrics

3. PROBLEMS ENCOUNTERED

- need to abandon spatial attention method because the training is too slow

4. SELF EVALUATION OF THE PROGRESS

- on track



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3 Year 3	Study week no.: 4 - 6
Student Name & ID: Gan Kian Yu & 19ACB01693	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- generate a web-based demo
- apply contrastive regularization with different distance metrics

2. WORK TO BE DONE

- analyze results per crime category

3. PROBLEMS ENCOUNTERED

- ShanghaiTech dataset not found (video corrupted / test video not given)

4. SELF EVALUATION OF THE PROGRESS

- on track



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3 Year 3	Study week no.: 6 - 8
Student Name & ID: Gan Kian Yu & 19ACB01693	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

-analyze results per crime category

2. WORK TO BE DONE

-Understand and implement the new VAD (MGFN, Y. Chen et al., 2022)

3. PROBLEMS ENCOUNTERED

-no problem encountered

4. SELF EVALUATION OF THE PROGRESS

-on track



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3 Year 3	Study week no.: 8 - 10
Student Name & ID: Gan Kian Yu & 19ACB01693	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Understand and implement the new VAD (MGFN, Y. Chen et al., 2022)

2. WORK TO BE DONE

-propose new method on VAD

3. PROBLEMS ENCOUNTERED

- not able to get the performance as high as the result stated in the paper of MGFN

4. SELF EVALUATION OF THE PROGRESS

-need to read more papers to propose the new idea and further improve the model performance



Supervisor's signature



Student's signature

POSTER

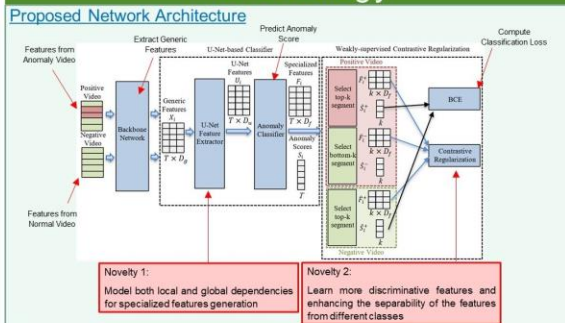
Abstract

Video anomaly detection (VAD) which is able to automatically identify the location of the anomaly event that happened in the video is one of the current hot study areas in deep learning. Due to expensive frame-level annotation in video samples, most of the VAD are trained with the **weakly-supervised method**. Temporal dependencies are critical to detect anomaly events. However, none of the previous works has modelled both local and global temporal dependencies effectively. In this project, we propose to use **U-Net** like structure to model both local and global dependencies for specialized features generation. Previous works has applied data augmentation, noise injection, l2 normalization, dropout regularization and special heuristics to prevent overfitting. However, none of the existing work have extended the feature-based approach to regularization where the strategy is to learn more discriminative features. In this aspect, we extend **contrastive regularization** in weakly-supervised manner as a new regularization technique to reduce overfitting. Experimental results show that our model achieves the second highest AUC performance compared to all published work.

Problem Statement & Objective

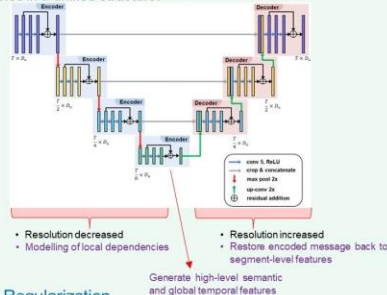
	Existing Method	Issue	Proposed Method
Temporal Modelling	<ul style="list-style-type: none"> RNN (Luo et al., 2017), temporal consistency (Liu, et al., 2018) and ConvLSTM (Liu et al., 2019) GCN-based methods (Wu et al., 2020) RTFM (Tian et al., 2021) 	<ul style="list-style-type: none"> can only capture short range dependencies slower and more difficult to train short and long temporal dependencies are captured separately 	<ul style="list-style-type: none"> U-Net structure to model both local and global dependencies for specialized features generation.
Regularization Techniques to reduce Overfitting	<ul style="list-style-type: none"> Decrease complexity of the network Data augmentation, noise injection, l2 normalization, dropout regularization 	<ul style="list-style-type: none"> None of the regularization technique focus on learning more discriminative features 	<ul style="list-style-type: none"> Contrastive regularization to enhance separability between normal and abnormal features makes the network less vulnerable to overfitting

Methodology



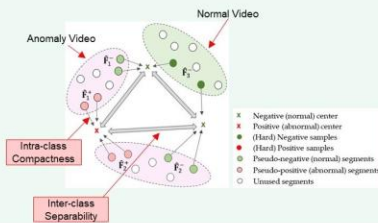
U-Net Temporal Modelling

- The U-Net network is novel used to model both local and global temporal dependencies in a unified structure.



Contrastive Regularization

- Contrastive regularization reduces overfitting by creating more robust features through intra-class compactness and inter-class separability.



Results

Dataset

	UCF-Crime
Total Number of Sample	<ul style="list-style-type: none"> 1610 training videos 290 testing videos
Annotation	<ul style="list-style-type: none"> Video-level labels for training videos Frame-level labels for testing videos
Duration of Video Sample	<ul style="list-style-type: none"> 8 seconds to 9 hours
Characteristic	<ul style="list-style-type: none"> Challenging due to diverse and complex background

Comparison with State-of-the-art Methods on UCF-Crime

- Our proposed method achieves state-of-the-art performance compared to recently published work.

Table 1. Comparison of AUC performance with other state-of-the-art methods on UCF-Crime

Supervision	Method	Feature	AUC (%)
Unsupervised	Sparsity-Combination (Li et al., 2013)	C3D RGB	65.51
	BODS (Wang et al., 2019)	13D RGB	68.26
	GODS (Wang et al., 2019)	13D RGB	70.46
	MIL-ranking (Sultani et al., 2019)	C3D RGB	75.41
	MIST (Feng et al., 2021)	C3D RGB	81.40
Weakly supervised	RTFM (Tian et al., 2021)	C3D RGB	83.28
	MIL-ranking (Sultani et al., 2019)	13D RGB	77.92
	MIST (Feng et al., 2021)	13D RGB	82.30
	RTFM (Tian et al., 2021)	13D RGB	84.30
	HCE (Lv et al., 2021) (original training dataset)	13D RGB	84.44
	HCE (Lv et al., 2021) (noise-augmented dataset)	13D RGB	85.38
	Ours	13D RGB	85.24

The best performing model HCE trains on a noise-augmented dataset. Without noise augmentation, our proposed model outperforms HCE.

Visualization of Model Performance


- The anomaly scores generated by our network is well aligned with the ground truth. Our model is able to localize the anomalous segments in the video.



Conclusion

- U-net is able to model local and global temporal dependencies more effectively.
- Contrastive regularization is able to generate more robust features that can further prevent overfitting.

PLAGIARISM CHECK RESULT



Originality Report

Processed on: 24-Apr-2023 21:25 +08
 ID: 2073983578
 Word Count: 7953
 Submitted: 1

FYP2_GanKianYu

By Kian Yu Gan

Similarity by Source	
Similarity Index	12%
Internet Sources:	4%
Publications:	11%
Student Papers:	1%

Document Viewer

include quoted | include bibliography | excluding matches < 8 words | mode: show highest matches together

Video anomaly detection (VAD) which is able to automatically identify

the location of the anomaly event that happened in the video

22

is one of the current hot study areas in deep learning. Due to expensive frame-level annotation in video samples, most of the VAD are trained with the weakly-supervised method. In

a weakly-supervised manner, the labels are at video level

37

. VAD is still an open question and challenging task because the model is trained with a limited sample in weakly supervised video-level labels. In this project, we aim to improve the VAD network with 2 different aspects. Firstly, we explore a technique to model the

local and global temporal dependencies. Temporal dependencies are critical to

10

detect anomaly events. Previous methods such as stacked RNN, temporal consistency and ConvLSTM can only capture short-range dependencies. GCN-based methods can model long-range dependencies, but they are slower and more difficult to train. RTFM captures both the short and long-temporal dependencies using two parallel structures, one for each type. However, the two dependencies are considered separately, neglecting the close relationship between them. In this aspect, we propose to use U-Net like structure to model both local and global dependencies for specialized features generation. Second, we explore a new regularization technique in a weakly-supervised manner to reduce overfitting. Insufficient training samples will lead to overfitting easily. Generally, the overfitting issue can be improved by reducing the complexity of the network, data augmentation, injecting noise into the network or applying dropout regularization. For VAD, previous works have applied special heuristics such as sparsity constraint and temporal smoothness to regulate the output of the model. However, none of the existing work has extended a feature-based approach to regularization where the strategy is

to learn more discriminative features. In this project, we extend contrastive

3

1 1% match ()
[Tian, Yu, Pang, Guansong, Chen, Yuanhong, Singh, Rajvinder, Verjans, Johan W., Carneiro, Gustavo. "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning", 2021](#)

2 1% match ()
[Feng, Jia-Chang, Hong, Fa-Ting, Zheng, Wei-Shi. "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection", 2021](#)

3 1% match ("Medical Image Computing and Computer Assisted Intervention – MICCAI 2021", Springer Science and Business Media LLC, 2021)
["Medical Image Computing and Computer Assisted Intervention – MICCAI 2021", Springer Science and Business Media LLC, 2021](#)

4 < 1% match ()
[Davi D. de Paula, Denis H. P. Salvadeo, Darlan M. N. de Araujo. "CamNuvem: A Robbery Dataset for Video Anomaly Detection", Sensors \(Basel, Switzerland\)](#)

5 < 1% match ()
[H. Soumare, S. Rezgui, N. Gmatj, A. Benkahla. "New neural network classification method for individuals ancestry prediction from SNPs data", BioData Mining](#)

6 < 1% match ()
[Eunsan Jo, MyoungHo Sunwoo, Minchul Lee. "Vehicle Trajectory Prediction Using Hierarchical Graph Neural Network for Considering Interaction among Multimodal Maneuvers", Sensors \(Basel, Switzerland\)](#)

7 < 1% match (Internet from 22-Nov-2022)
<https://deepai.org/publication/mist-multiple-instance-self-training-framework-for-video->

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	GAN KIAN YU
ID Number(s)	19ACB01693
Programme / Course	Bachelor of Computer Science (Honours)
Title of Final Year Project	Video Anomaly Detection with U-Net Temporal Modelling and Contrastive Regularization

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: <u>12</u> % Similarity by source Internet Sources: <u>4</u> % Publications: <u>11</u> % Student Papers: <u>1</u> %	
Number of individual sources listed of more than 3% similarity: <u>0</u>	
Parameters of originality required and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: Tan Hung Khooon

Date: 26/04/2023

-

Signature of Co-Supervisor

Name: -

Date: -

Bachelor of Computer Science (Honours)
Faculty of Information and Communication Technology (Kampar Campus), UTAR



UNIVERSITI TUNKU ABDUL RAHMAN

**FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY
(KAMPAR CAMPUS)**

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	19ACB01693
Student Name	Gan Kian Yu
Supervisor Name	Ts Dr Tan Hung Khoon

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
	Front Plastic Cover (for hardcopy)
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)

Date: 24/4/2023