COLLABORATIVE BATCH LEARNING FOR CRIME SCENE DETECTION

BY

TOH YUE XIANG

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 2022

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: COLLABORATIVE BATCH LEARNING FOR CRIME SCENE DETECTION

Academic Session: JANUARY 2022

I TOH YUE XIANG

(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:
9, JALAN SONGKIT 2,
TAMAN SENTOSA 80150
JOHOR BAHRU, JOHOR

Tan Hung Khoon
Supervisor's name

Date: 19 April 2022

Date: 19/4/2022

FACULTY OF _INFORMATION AND COMMUNITCATION TECHNOLOGY_

UNIVERSITI TUNKU ABDUL RAHMAN

Date: _19 April 2022____

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that_____TOH YUE XIANG_____(ID No: __*18ACB01082*) has completed this final year project entitled "_COLLABORATIVE BATCH LEARNING FOR CRIME SCENE DETECTION_" 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 / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

_____
TOH YUE XIANG

*Delete whichever not applicable

DECLARATION OF ORIGINALITY

I declare that this report entitled "**COLLABORATIVE BATCH LEARNING FOR CRIME SCENE DETECTION**" 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 : _TOH YUE XIANG_____

Date : _19 April 2022_____

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisors, Dr Tan Hung Khoon who has given me this bright opportunity to engage in a Computer Vision project. It is my first step to establish a career in Computer Vision field. A million thanks to you.

Then gratitude to a very special person in my life, Sarah, for her patience, unconditional support, and love, and for standing by my side during hard times. Finally, I must say thanks to my parents and my family for their love, support, and continuous encouragement throughout the course.

ABSTRACT

Surveillance camera is used in many settings to capture the real-life happenings. Lack of intelligent surveillance camera system decrease the effectiveness of surveillance camera in reducing crime. Our project developed a system to automatically detect crime scene event from the surveillance camera. In our project, we trained our model with normal and crime video from UCF crime dataset. Our work used I3D model pretrained on kinesis dataset to extract the feature frame by frame. We added an 1D dependency capturing attention module on top of the feature extractor to make the features extracted more useful and suitable for the dataset we were using. We used Multiple Instance Learning network as the framework of our system. Since, it was a weakly supervised learning model, the dataset that we used to train our model is weakly labelled dataset, this means that our dataset will not consist of the exact temporal segment where the anomalies happened in the surveillance video. Ranking loss function with sparsity and temporal smoothness constraint was used as our loss function to better detect the anomaly segment throughout the surveillance video.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Title Page

LIST OF ABBREVIATIONS

C3D	3D Convolutional Network
CNN	Convolutional Neural Network
FC	Fully Connected
MIL	Multiple Instance Learning
R3D	3D Residual Network
C3D	3D Convolutional Network
CNN	Convolutional Neural Network
Spatiotemporal	Spatial and Temporal
ConvNet	Convolution Network

CHAPTER 1

Introduction

Surveillance camera is treated as a security measure in many countries to reduce the crime rate index. According to research, there are about 770 million surveillance cameras in use now and this figure is expected to increase to 1 billion by the end of 2021 [1]. The number of people per surveillance camera in China increase from 1 camera for 4.1 people in 2018 to 1 camera for 3.37 people in 2019 [2].However, it is found that the crime rate index is not correlated to the number of surveillance camera used. One of the reasons to this is because the lack of artificial intelligence (AI) technology adaptation in the surveillance camera. The most used AI technology in surveillance currently is face recognition system. It is helpful in identifying the criminal identity only which is for solving crime case, but not in the identifying the crime event which is for crime prevention. Hence, an intelligent surveillance camera system with anomaly detection system is needed to reduce crime rate index.

Anomaly detection system is a system to detect crime scene segment in the surveillance camera. For example, we input the surveillance footage, then the system will output the start and end time where the crime scene lies within the footage and what crime class the temporal segment belongs to. However, developing anomaly detection system is a challenging task in computer vision due to the absence of large fully annotated crime video dataset. Also, it is hard to judge an event as anomaly or normal sometimes. Hence, many researches have been ongoing to develop good performing crime scene detection system.

Some of the earlier work developed a specific type of anomaly detection system such as violence detection system [3]. Some other work exploited on normal video because normal videos were available in larger quantity. For instance, deep auto-encoder based approach by [4] and dictionary-based approach by [5].A more recent work proposed to use Multiple Instance Learning (MIL) model to develop an anomaly detection system [6]. Their work exploited weakly labelled normal and abnormal video dataset and achieve significant performance. Their proposed approach was set up in weakly supervised setting which meant that only video level annotated dataset was needed to train their network.

Our project was like [6] work in which we exploited on both normal and anomaly videos and train using MIL model. However, we changed the feature extraction network to a deeper 3D network which was I3D network. We also proposed to add an attention module on top of the feature extraction network to extract the feature on region level. By modifying the feature extraction network, we believed that the performance of our system would increase since more descriptive and robust features were used to train the MIL model compared to the work in [6].

1.1 Project Scope and Objectives

The main objective of our project was to develop an improved version of anomaly detection system with MIL model which can yield higher accuracy. Our reimplementation of MIL model with exact same setting as mentioned in [6] has accuracy of 70.58%. We aim to increase this accuracy with deeper feature extraction network and including an attention module. To achieve this main objective, we did the following:

1. To develop an attention module to extract more refined feature.

This attention module was developed on top of the I3D network to assign weight distribution to the feature extracted from I3D network. The features output by the attention module was region level feature where it focused on the specific feature of the whole feature.

2. To develop a crime scene detection web app.

A web app was developed using our trained model to predict the surveillance footage. The input video was passed into the crime scene detection system which consist of feature extraction network, attention module and MIL model. Then the system output the prediction score for each segment in the input video.

1.2 Report Organization

The details of this research are shown in the following chapters. In Chapter 2, we discussed some related study that we have reviewed. Then, we discussed the methodology of our proposed approach in detail in Chapter 3. Then, Chapter 4 described the detailed implementation of our model and also the experiment setup. In Chapter 5, we discussed the evaluation result of our model. Lastly, we summarized our findings and give recommendation for future work in chapter 5.

CHAPTER 2

Literature Review

We separated this chapter into 3 sections. In section 2.1, we reviewed on some background study on the action recognition such as C3D, I3D and R(2+1)D. In section 2.2, we reviewed on weakly supervised anomaly detection system. In section 2.3, we reviewed the work on attention module.

2.1 Action recognition

The breakthrough of AlexNet in image recognition task using 2D deep learning framework had opened new era of computer vision. Some researches had been done using 2D deep learning framework to try push video analysis task to achieve the same breakthrough AlexNet had. However, 2D network cannot work well in video analysis simply because temporal dimension is not processed in 2D network. As shown in figure 2.1.1(a), temporal dimension collapsed after passing through 2D convolution, even when the 2D convolution convolved on multiple frames, 3D information still collapsed into 2D. Hence, it was found that 2D convolution was not suitable for video analysis task, no matter how deep it was.

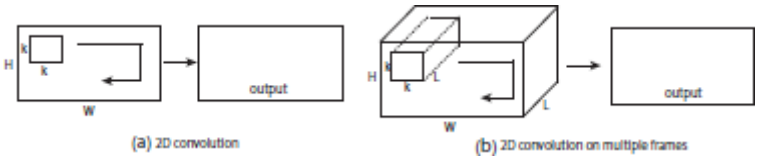


Figure 2.1.1 (a) 2D CNN (b) 3D CNN [7]

To solve this problem, [7]experiment on 3D convolution. In their research, they passed the video input to a 3D convolution and found out that the temporal dimension of the input was kept as shown in figure 2.1.1(b). This solved the problem caused by 2D convolution. After realising that 3D reserve the temporal signal, they built a C3D network to perform action recognition task as shown in figure 2.1.2. The network was made up of eight 3D convolutional layer, five 3D max-pooling layer and 2 fully connected layer with 4096 neurons. Each of the convolutional layer had 3x3x3 kernel with stride 1 which covered spatiotemporal dimensions. The kernel size of each convolutional layer is as shown in figure 2.1.2

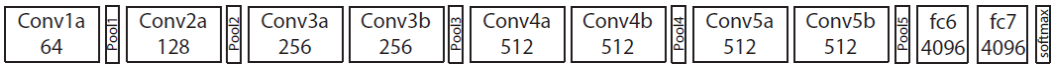


Figure 2.1.2 C3D network architecture [7]

The accuracy of their proposed method significantly increased and outperformed the baseline model iDT [8] with Fisher vector and Imagenet [9]. This had proven that temporal dimension is not to be neglected for video analysis task. It opened new possibility in video analysis task. However, the C3D network proposed in this paper was not deep since it was only made up by 8 layers. The breakthrough like AlexNet was not seen yet.

I3D network was similar to C3D network, but it went down further to extract the feature on single spatial RGB stream and temporal stream which is the motion flow of moving objects between consecutive frame along the vertical and horizontal axis as shown in figure 2.1.3.

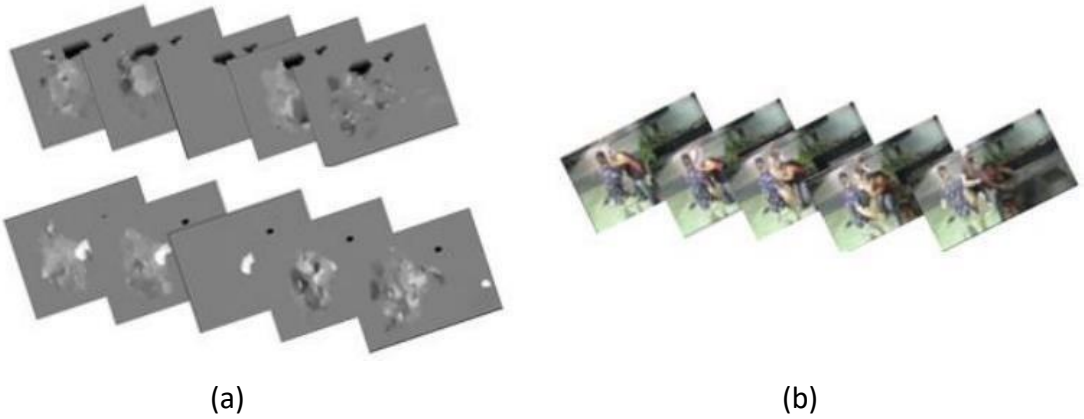


Figure 2.1.3 A pair of optical frames generated (a) based on the consecutive RGB frames (b) [10]

Despite single RGB stream network is sufficient in action recognition as shown in the work by [7], including temporal stream improved the accuracy result even more. As shown in figure 2.1.4, the I3D model proposed by [10] used pretrained 2D ConvNet as the feature extractor, then inflated the feature into 3D with global pooling layer before pumping into pretrained C3D model.

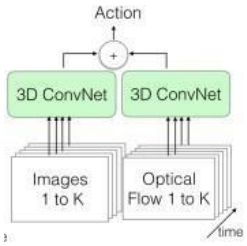


Figure 2.1.4 I3D model architecture

I3D not only has way lesser parameter and better accuracy compare to C3D model, it is also capable of recognizing many types of activity and salient action since it was pretrained on Kinetics Human Action Video dataset, which has 400 human-action classes, which each consists of 400 video clips.

Another work [11] that realized C3D might be too shallow, took further step in experimenting much deeper 3D network. They experimented on multiple deep learning frameworks such as R2D, mixed convolution network, reversed mixed convolution network, R3D and R(2+1)D. The R3D and R(2+1)D was further separated into 2 types which are 18 layers and 34 layers. R(2+1)D was a special type of network that does not violate 3D convolution principle. The 3D convolution was factorized into (2+1)D convolution in R(2+1)D network as shown in figure 2.1.3

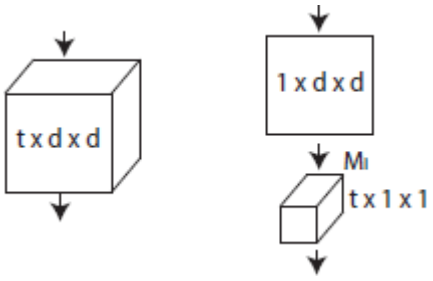


Figure 2.1.5 R(2+1)D [11]

Without questions, R3D and R(2+1)D outperformed all the other network in their paper, because these 2 networks used 3D convolution. Another interesting findings in their paper was that R(2+1)D network perform slightly better than R3D network while requiring lesser computation power because it consisted of slightly lesser parameter. The proposed R(2+1)D outperformed state-of-the-art methods such as DeepVideo [12], C3D [7] and P3D [13]. Hence, in this paper they had developed a deeper 3D convolutional network which yield comparable result with the state-of-the-art method. They had also proven that deeper network can increase

the performance of the network and confirmed again that temporal signal is important in video analysis task.

2.2 Anomaly detection

Anomaly detection system is an important factor to develop an intelligent surveillance camera so that crime rate index can be reduced effectively. However, it is a challenging task. One of a reason could be the absence of fully annotated surveillance video dataset for many types of anomaly events is not available in large scale. Several attempts were proposed to solve this problem, that was to develop a specific type of anomaly detection system such as violence detector and traffic detector [3], [14] and [15]. However, this type of system could not be generalized into another anomaly class which was not very useful. Hence, these types of anomaly detection system were not realistic, because it was too problem specific and could not be applied on many other types of anomaly event.

To develop a more generalized anomaly detection system, some researchers focus to develop detector that train only with normal video dataset, because normal video was available in large scale. In their theory, any motion that deviate from these normal video datasets will be classified as anomaly [16] and [17]. However, it was difficult to obtain all normal motion for training. Hence, some works proposed to learn global motion pattern. For instance, topic modelling [18], Hidden Markov Model (HMM) on local spatio-temporal volumes [19], motion patterns [20], social force models [21], histogram-based methods [22], context-driven methods [23] and mixtures of dynamic textures model [24].

Some works propose to use sparse representation and dictionary learning approaches to develop anomaly detector through learning normal behaviour, because of the success of this technique in other computer vision domain [25] and [5]. In their theory, the network classified the video as anomalous if the pattern yields high reconstruction errors. Inspired by the success of deep learning in image classification, some approaches adapted deep learning method into anomaly detection. For instance, deep learning based autoencoders by [26] and [4] to learn normal behaviour. However, this type of detectors which only learn normal behaviour has high false positive rate, because it was impossible to learn all type of normal pattern. Besides, it was hard to define normal event since normal event can varies person to person too.

A recent work by [6] proposed Multiple Instance Learning (MIL) network as shown in figure 2.2.1 solved the problem of anomaly detection system mentioned previously. This network was developed in weakly supervised manner. This means it did not need to train on fully annotated dataset. The dataset passed to the network was only video level annotated which means the network only know what action class the video belongs to, but it did not know the start time and end time of the crime scene within the video. Besides, their proposed network was trained with both normal and abnormal dataset. Due to this, their network was able to produce lower false alarm rate for normal video. Hence their proposed network solved the problem of lack of fully annotated training dataset and solve the problem caused by training only with normal videos.

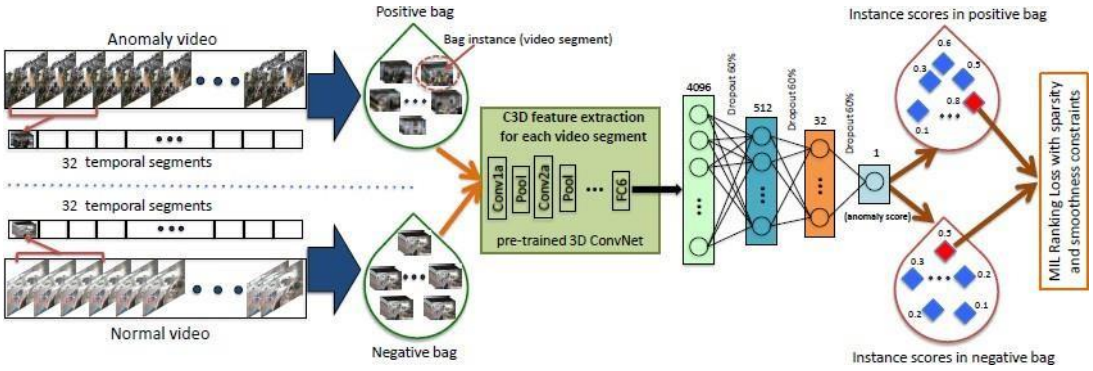


Figure 2.2.1 MIL model architecture [6]

Another interesting finding in their proposed model was the used of ranking loss instead of usual classification loss as the loss function to train their network. They trained their model in a regression manner. This was because the dataset they used did not contain any ground truth. Besides, there was no clear cut between anomaly video and normal video. Hence, they chose a ranking loss to rank the segment which likely the segment belonged to.

Their proposed network outperformed the baseline model such as Binary classifier, dictionary-based approach [5] and deep learning autoencoder [4] with significantly higher accuracy and lower false alarm rate. Hence, they opened a new dimension for weakly supervised setting in anomaly detection system. However, the performance of their proposed method was still far from perfection.

One of the reasons could be the feature they were using to train their MIL model. In their proposed approach, they extracted the feature from every 16 frame. This meant that the features

were extracted on frame-level only. This type of feature was not discriminative and representative enough for the MIL model to train. To solve this problem, [27] proposed to extract the feature on region level.

In their proposed approach, they separated the feature extraction module into 2 branches – interactive dynamic branch and spatial-temporal branch. In interactive dynamic branch, they incorporated interaction modelling to model the interaction between each region level feature to further gave the meaning to each region level feature. The video input will be pumped into a social force module to compute the social force map first. This social force map will reflect the dynamic interaction and the degree of the interaction in the frame. Then this social force map would be passed into C3D feature extraction network to extract the 3D feature from these maps.

In both branches, they added attention module on top of the C3D feature extraction model. Their attention module was trained on both spatial dimension and temporal dimension. This attention module assigned different weight distribution to each feature extracted. By doing this, their network could identify the important part which had more weight assigned and neglect the potential background which has less weight assigned on the feature vector. Then these 2 features extraction module would be merged together before pumping into the MIL model which has the same setting as proposed in [6]. Their proposed network outperformed the MIL model which only extract feature on frame-level. This had proven that extracting region-level feature would be useful for MIL model.

2.3 Attention module

The attention module used in [27] is 1-D dependency capturing attention module as shown in figure 2.3.1. The 2048 1D feature would be passed to 3 different 1D convolution layer to produce feature vector B, C and D. The spatial attention weight was generated by matrix multiplication of feature vector B and C. Then this spatial attention weight would undergo matrix multiplication with feature vector D. The final feature vector with attention weight distribution is produced by concatenating original 1D feature vector with the final feature vector D. This 1-D dependency capturing attention module allowed us to capture the global dependency in the original feature vector.

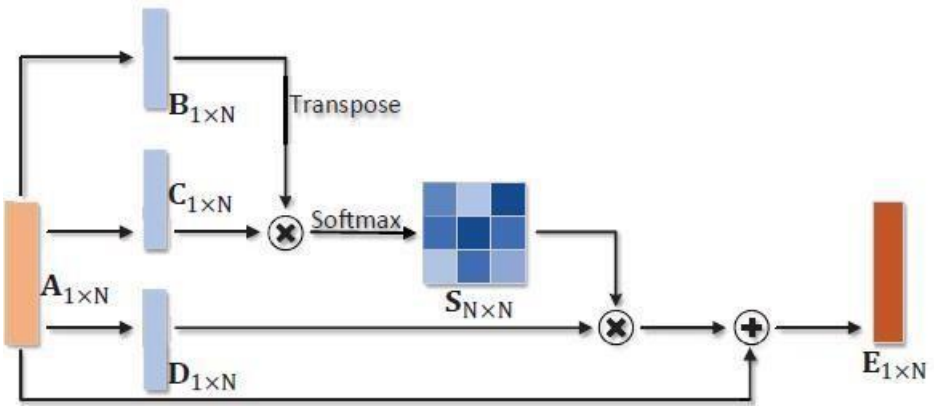


Figure 2.3.1 1D dependency attention capturing module [27]

Although the implementation of 1-D dependency capturing attention module improve the performance of anomaly detection task, there is an underlying issue in the attention module. The spatial attention weight was forged by matrix multiplication of 2 1D feature vector. It was not output from the true heigh and length dimension of the individual video segment.

Hence in another approach of attention module, [28] proposed a video vision transformer to capture the attention from video. In their paper, they proposed a new way to perform positional embedding and several variants of pure transformer architecture. Their proposed positional embedding method, tubelet embedding as shown in figure 2.3.3 showed improvement in accuracy compare to the normal positional embedding method, uniform frame embedding as shown in figure 2.3.2. Uniform frame embedding was to extract uniform fragment from individual segment, then embed temporal segment by temporal segment into a linear feature vector. Whereas tubelet embedding method was to extract non-overlapping, spatio-temporal fragment from the input, then embed it linearly into feature vector. By doing this, the tokens were embedded from temporal, height and width dimensions respectively which was different from uniform frame embedding where the temporal information were fused by the transformer.

Rd

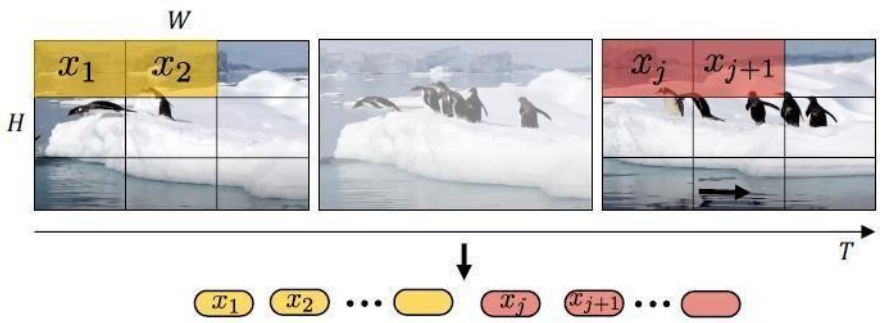


Figure 2.3.2 Uniform frame embedding [28]

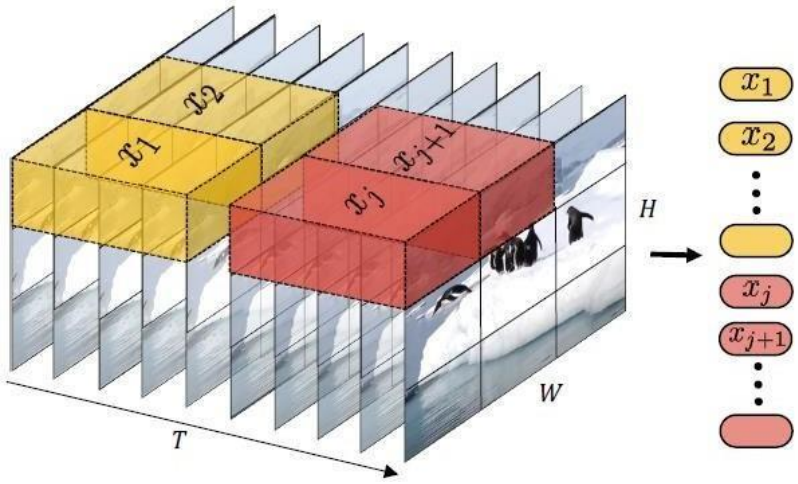


Figure 2.3.3 Tubelet Embedding [28]

In their paper, they proposed 4 variants of transformer-based attention module. Firstly, spatio-temporal attention where the tokens were all pumped into transformer encoder to model pairwise interaction which was also known as Multi-Headed Self Attention (MSA). However, the complexity of this attention module was quadratic with respect to the number of tokens. Second model was factorized encoder which split the transformer encoder into temporal and spatial as shown in figure 2.3.4. The spatial encoder modelled the interaction between the token with the same temporal index, whereas the temporal encoder modelled the interaction between the tokens from different temporal indices. The output from transformer encoder was then pumped into the classification layer. The floating point (FLOPs) of this model was lesser compare to the spatio-temporal attention but it has more transformer layers.

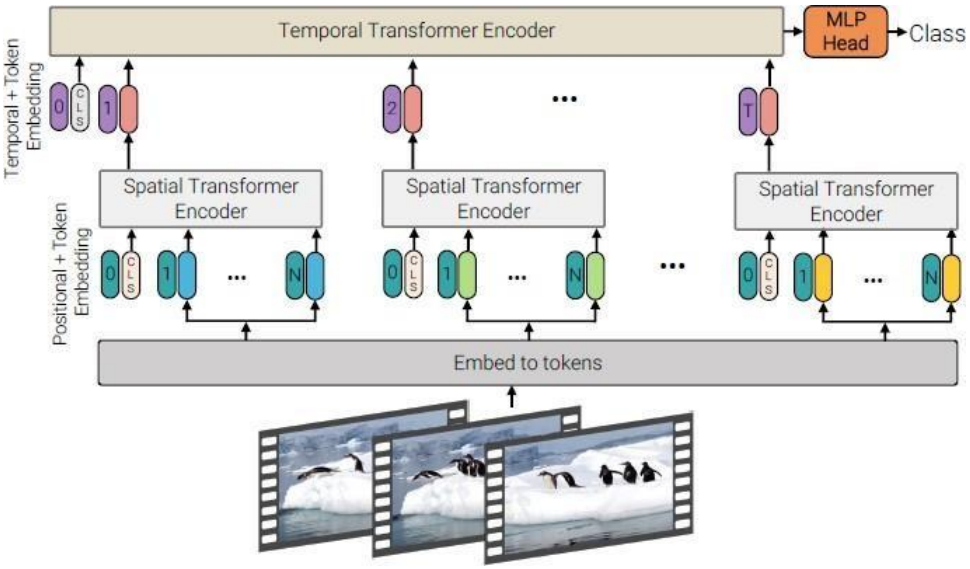


Figure 2.3.4 Factorized Encoder [28]

The third model was factorized self-attention as shown in figure 2.3.5 where the number of transformer layer was the same as spatio-temporal attention. The self-attention was computed spatially then temporally or temporally then spatially. Due to factorization of self-attention computation this model was more efficient than the spatio-temporal attention which same number of transformer layer and achieved same computation complexity as factorized encoder but with lesser number of transformer layers. However, the number of parameters increased due to the additional self-attention layer.

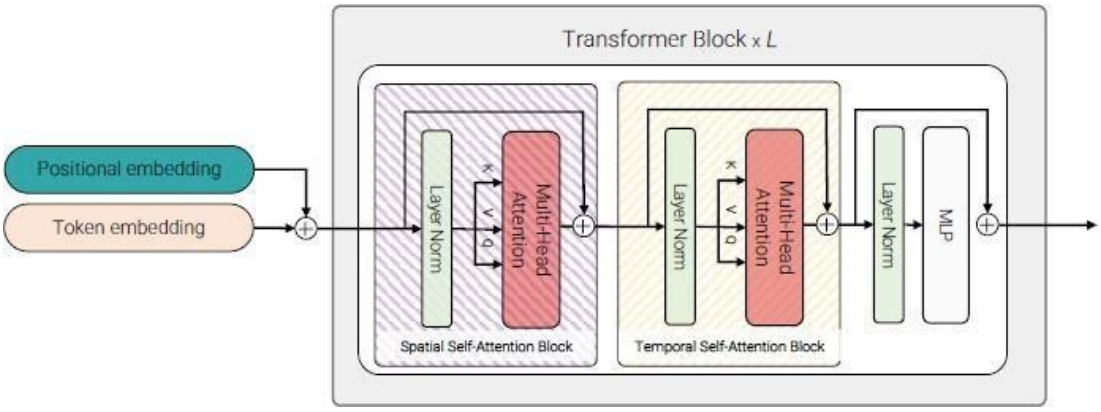


Figure 2.3.5 Factorized self-attention [28]

The last model was factorized dot-product attention where the attention for each token was computed separately over spatial and temporal dimension using different muti-head dot-product attention as shown in figure 2.3.6. This model achieved the same computational

complexity as factorized encoder and factorized self-attention, while retaining same number of parameters as spatio-temporal attention.

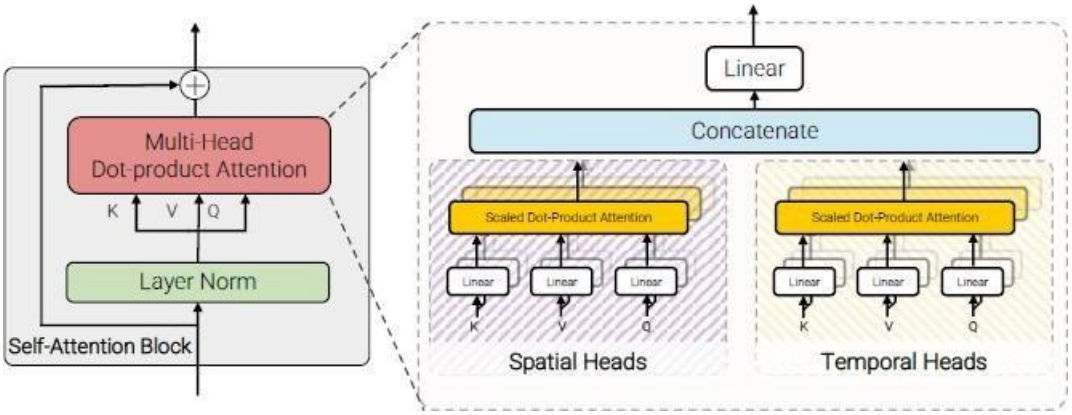


Figure 2.3.6 Factorized dot-product attention [28]

Among all the attention model, spatio-temporal attention had the best accuracy but require longest runtime. The factorized encoder had only slightly lower accuracy but it had the shortest runtime among the attention model. However, all the attention models improved the accuracy from the baseline.

Our project was similar to the approach proposed by [6]. We used the same dataset and MIL model proposed in their paper. However, we changed the C3D feature extractor into a deeper network which was I3D network and added an attention module on top of the feature extractor to assign different weight to different feature extracted to achieve region and temporal level feature extraction.

CHAPTER 3

Methodology

3.1 System Overview

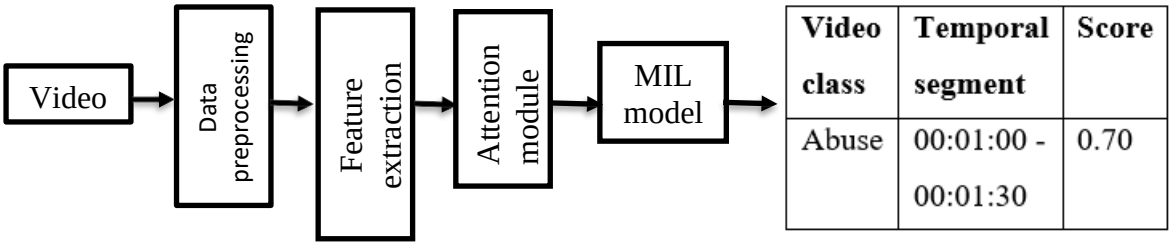


Figure 3.1.1 System design

We first segregated the surveillance footage into 32 segments. Then the segregated clip would be pumped into the feature extractor which was I3D network to output frame-level 1D feature vector. Then this feature vector would be simply forwarded to the attention module to output region level features. The region level features were then passed into the MIL model for classification and the score for each segment were predicted.

3.1.1 Data preprocessing

In data preprocessing, the long untrimmed surveillance video was segregated into 32 non overlapping, equal temporal segments. For MIL model, the videos were treated as a bag and the segregated temporal segment were the instances in the bag. The bags were categorized into 2 categories – Positive bag and negative bag. Positive bag consists of at least 1 anomaly segment whereas negative bag has no anomaly segment at all. Each instance was passed into the model one by one to predict anomaly score for itself.

3.1.2 Feature extraction

We selected I3D network as our feature extractor in our project, because I3D was a deeper and more robust network compare to C3D. The features were extracted from the last global average pooling layer which output 1024 features. Since the I3D network was pretrained on

kinetics dataset instead of UCF crime dataset that we were using. Hence, the feature extracted from UCF crime dataset might not be representative and discriminative enough.

3.1.3 Attention Module

According to [27], assigning different weight to different feature could improve the accuracy of action recognition. This was because the feature was now extracted on the lower level. Heavier weight was assigned to more important feature and lighter weights were assigned to the background features which were the noise and unimportant part to the training. For example, in a robbery scene, we wanted our network to focus on the robbers and the victim instead of the surrounding shown in figure 3.1.2. Then throughout the whole videos, we wanted the model to focus on the temporal segment when the robbery happened as shown in figure 3.1.3. In this case, more weight should be assigned to the feature representing the red box region, since it was the important part to describe the event. Hence, we added 1-D dependency capturing attention module on top of our feature extractor to achieve region level feature extraction.

Figure 3.1.2 Robbery Scene

Figure 3.1.3 Segments of robbery video

We simply forwarded the 1024 1D feature vector extracted from I3D model into 3 different 1D convolution layer to output 3 new 1D feature vector with same dimensions, f_1, f_2 and f_3. The affinity matrix, A was generated by forwarding the dot product of transposed f_1 with f_2 to a softmax layer as shown in the equation below:

$$A_{i,j} = \frac{\exp\,(f_{1i}^{T}.\, f_{2i})}{\sum_{i=1}^{N} \exp\,(f_{1i}^{T}.\, f_{2i})}$$

where $A_{i,j}$ was the representation of degree to which i^{th} position value affects j^{th} position value. The higher the correlation between i^{th} and j^{th} position the larger the value of $A_{i,j}$ in the affinity matrix.

Then the affinity matrix would be multiplied with f_3, before concatenating with the original feature vector as shown below:

$$weighted\ feature, f = f_{ori} + \beta(S * f_3)$$

where β was a scaled parameter to be trained and it was initialized to a very small number. The attention weights were assigned by concatenating the original feature vector with value produced by $\beta(S * f_3)$.

Different weight distributions were assigned by this 1-D dependency self-attention module, causing the value of all position inside feature vector to be correlated. We achieved region level feature extraction by capturing global dependencies that interrelated the distant features in the feature vector.

3.1.4 Anomaly Detection

The Multiple Instance Learning (MIL) model was trained in weakly supervised manner. This meant that our dataset did not need to have ground truth annotation which tells the network the exact start and end frame of the crime scene segment within the long untrimmed surveillance video. The network only required the action class label of each input. For example, the video was from abuse class, arrest class, or normal class. The reason we adapted weakly

supervised learning was because there was no fully annotated dataset available in large scale for training.

The MIL model would also be trained with both anomalous and normal segment, so that the network could learn better on different action class. This was because every human behaves differently in normal and anomalous event. Providing the network both type of classes to learn, allow our model to learn more variation of human patterns. By doing so, we could also reduce the false alarm rate.

3.1.5 Ranking loss function

Ranking loss was a regression loss to train the network. Our project employed ranking loss because the dataset we were using was not fully annotated. Hence, it was easier for the model to rank the segment whether it was more likely to be anomaly segment or normal segment instead of distinctly classifying it as 1 or 0. We would discuss the detail of ranking loss function used in the following paragraph.

In our project, the loss function was based on the Support Vector Machine (SVM) hinge loss function as shown in equation below:

$$L = \sum_{j=1} \max\,(0,1 - f(V_t^i) + f(V_f^i))$$

where $f(V_t^i)$ was the predicted score for true class sample i and $f(V_f^i)$ was the predicted score of the non-true class sample i. However, it was hard to classify the video in 1/0 matter like a classification manner, because there was no clear cut between normal and abnormal video. Therefore, training in regression manner was easier and more suitable in our project. Due to this, we modified the SVM hinge loss function as below:

$$L = \max\,(0,1 - f(V_{pos}^i) + f(V_{neg}^i))$$

where $f(V_{pos}^i)$ denoted the score of instances i in the positive and $f(V_{neg}^i)$ denoted the score of instances i in the negative bag. In an ideal case, $f(V_{pos}^i)$ was expected to be larger than

$f(V^i_{neg})$ by margin of 1 to incur 0 loss. Else, penalty would be incurred. However, since the dataset we were using was weakly labelled dataset and we did not know exactly at which segment the crime scene occurred, so we could not implement $f(V^i_{pos}) > f(V^i_{neg})$. Hence, we needed to enforce the ranking with maximum score of each instance as below:

$$\max\,(f(V^i_{pos})) > \max\,(f(V^i_{neg}))$$

where the maximum of predicted score of all instances in the positive bag is enforced to be larger than the maximum of predicted score of all instances in the negative bag. The $\max\,(f(V^i_{abn}))$ was the instance that was very true positive whereas $\max\,(f(V^i_{nor}))$ was the instance that is mostly like to be false positive – normal segment that was classified as anomalous segment. Since the maximum score of positive instances was enforced to be higher than maximum score of negative instances, our loss function was still not violated. With all the rules and condition enforced, our MIL ranking loss function was formulated as:

$$l(B_{abn}, B_{nor}) = \max\,(0, 1 - \max_{i \in B_{abn}} (f(V^i_{abn})) + \max_{i \in B_{nor}} (f(V^i_{nor})))$$

The score output from our model was for every video segment in the video. To avoid the occurrence of huge gap between the score of contiguous segments, we needed to have temporal smoothness and sparsity constraint in our loss function. Therefore, with 2 more constraint introduces, our final loss function was formulated as below:

$$l(B_{abn}, B_{nor}) = \max\,(0, 1 - \max_{i \in B_{abn}} (f(V^i_{abn})) + \max_{i \in B_{nor}} (f(V^i_{nor})))$$
$$+ \lambda \sum_{i}^{(n-1)} (f(V^i_{abn}) - f(V^i_{nor}))^2 + \lambda \sum_{i}^{n} (f(V^i_{abn})) + ||W||$$

where $\lambda \sum_{i}^{(n-1)} (f(V^i_{abn}) - f(V^i_{nor}))^2$ was the temporal smoothness and $\lambda \sum_{i}^{n} (f(V^i_{abn}))$ was the sparsity constraint in n total number of segments. W represented the weight parameter of the model. The λ of temporal and smoothness constraint was the hyperparameter for the constraint. The batch loss computed from each run would be obtained by averaging over the

number of samples per batch which was 60 in our use case. The weight of the model was updated during backpropagation via the gradient computed.

CHAPTER 4

Implementation Detail

4.1 UCF Crime dataset

The dataset that we were using was UCF Crime video dataset. It consisted of 13 classes of anomaly video and 1 class of normal video. It was a weakly labelled video dataset with annotation at video-level only. This meant the network would not know exactly at which temporal segment the crime scene appeared. The network would only know which anomaly class this video belongs to. However, the testing split were fully annotated for evaluation purpose. The detail of each class was as shown below.

Class	**Train split**	**Test split**
Abuse	48	2
Arrest	45	5
Arson	41	9
Assault	47	3
Burglary	87	13
Explosion	29	21
Fighting	45	5

Class	**Train split**	**Test split**
Road accidents	127	23
Robbery	145	5
Shooting	27	23
Shoplifting	29	21
Stealing	95	5
Vandalism	45	5
Normal	800	150

Table 4.1.1 UCF Crime Dataset

4.2 Implementation details

The setting for videos input were 240 x 320 pixels and frame rate 30 frame per second. The features were extracted using pretrained I3D network with center crop at every 16 frames. Then we segregated the features extracted into 32 non overlapping temporal segments where each segment contained the average of 16-frames features. In our experiment, we built 3 different modules on top of I3D feature extractor to compare the difference. The model architecture for our first model, 1D dependency capturing attention module was shown in figure 4.2.1. We used ReLU activation in every 1D convolution layer. The kernel size of 1D convolution layers was set to 3. The setting of batch normalization layers was the tensorflow default setting. We set the number of head to 2 after performing hyperparameter tuning with different number of heads.

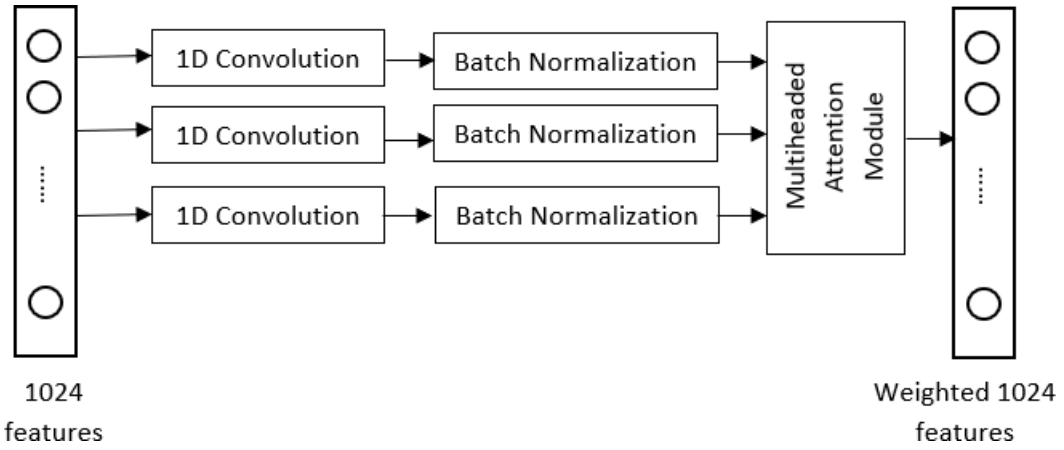


Figure 4.2.1 Block diagram of 1D dependency attention capturing module

For second type of modules as shown in figure 4.2.2, we forwarded the feature extracted into 2 different 1D convolution layer to compute the affinity matrix. Batch normalization layers were added after each 1D convolution layer. The setting of 1D convolution layer and batch normalization were the same as 1D dependency attention capturing module. Then we flatten the affinity matrix into 1D feature vector with 1024 features to match with the input dimension of the first layer of the MIL model.

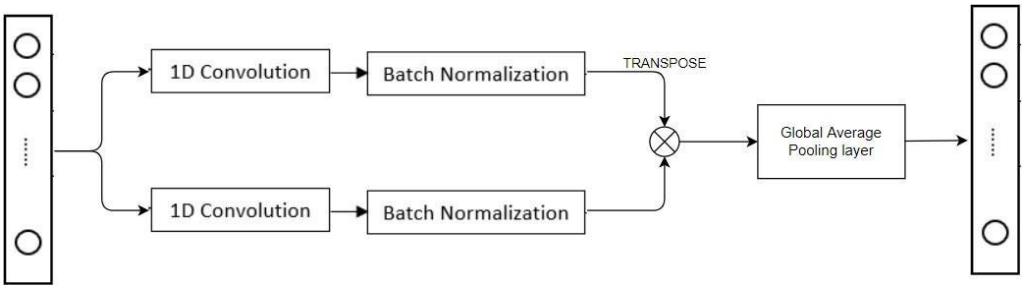


Figure 4.2.2 Block Diagram of affinity matrix module

For the last module as shown in figure 4.2.3, we simply forwarded the feature to a 1D convolution layer to increase the complexity of the feature. Batch normalization layer was added after the 1D convolution layer. The setting of 1D convolution layer and batch normalization were the same as 1D dependency attention capturing module as well.

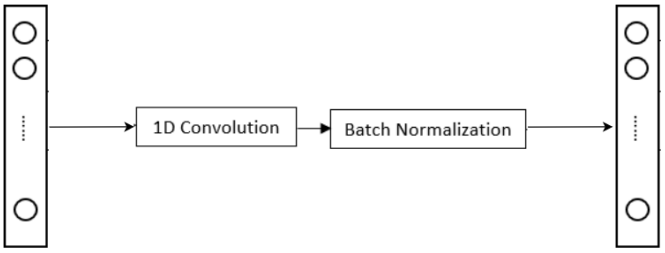


Figure 4.2.3 Block Diagram of increasing complexity module

The MIL model was made up of 3 fully-connected (FC) layer with 60% dropout layer after each FC layer. To further prevent overfitting issue, we also added L2 regularization which was set to 0.001 in all our FC layers. The first FC layer had 1024 neurons, followed by 128 neurons in the second layer, and 8 neurons in the final FC layer. We used ReLU activation for layer 1 and layer 3 only, and sigmoid activation for the final layer to predict the anomaly score for the instance. The full MIL model was shown in figure 4.2.2. Adagrad optimizer with learning rate set to 0.001 was used in our implementation. For the hyperparameter of temporal and smoothness constraint, we set it to 8 x 10^{-5} which is the same setting as [6].

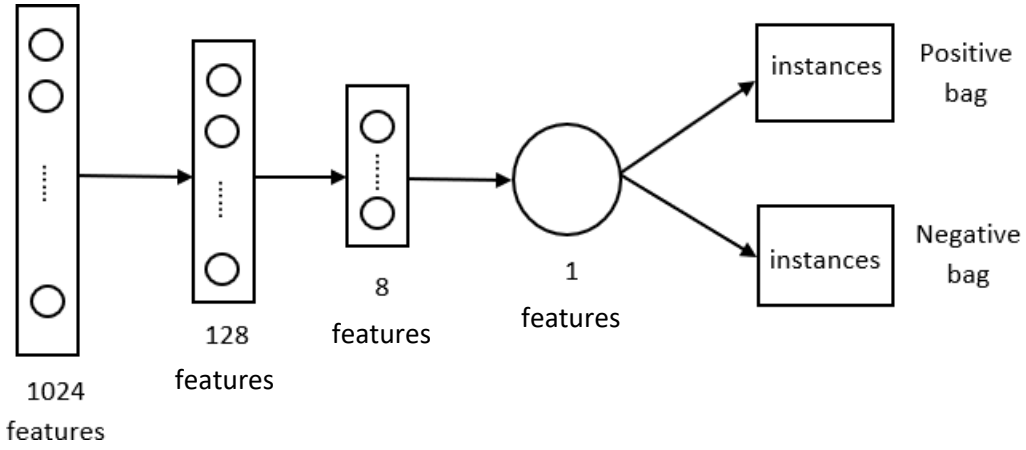


Figure 4.2.4 MIL model

The MIL model was trained with batch size of randomly selected 60 segregated video clips with 30 each from anomaly class and normal class per training. The gradient of the computational graph for each forward pass was computed with tensorflow backend. The score was computed for each instance which was the segregated temporal segments of the video. Then the loss for each batch was obtain with the ranking loss function from the backpropagation of these scores.

4.3 Test and evaluation

We used the model and weight obtained from our training to test and evaluate our model. Our model was evaluated with the test split from UCF crime dataset. The testing video were unseen videos by the training network. During testing, the predicted score for each testing video was distributed across the entire video after the temporal segment of the crime scene was determined. The ground truth temporal segment was available for testing video. This meant

that we knew the starting and ending frame of the crime scene in the video. Hence, we could calculate the AUC score with the ground truth and predicted score and plot the ROC for our model.

CHAPTER 5

Evaluation Result

5.1 Comparison between MIL with C3D feature extractor

We reimplemented the network proposed by [6] with the exact same setting using our system. As shown in table 5.1.1, the AUC score of the baseline model obtained was 0.7058. Our implementation using I3D feature extractor was 0.7252 which outperformed the baseline model. In the following section of report, we represented MIL model with I3D feature extractor as “our model”. This showed that using more robust and deeper feature extractor improved the performance of MIL model. We also implemented model with 1D dependency attention capturing module for both our model and baseline model. From the result shown in table 5.1.1, we observed that extracting the feature at region level further improved the performance significantly regardless of what feature extractor used.

Model	**C3D**	**I3D**
Without 1D dependency attention capturing module	0.7058	0.7252
With 1D dependency attention capturing module	0.7215	0.7476

Table 5.1.1 AUC Score of the models

We further experiment on our model with different type of module added on top of the feature extractor as shown in table 5.1.2. From the result shown in table 5.1.2, we observed that extracting the feature with increasing complexity module did not show much significant and for the model with affinity matrix module, the performance even decreased to lower than the baseline model. This was because the affinity matrix was flattened to form the 1D feature vector causing the affinity information obtained collapsed and lost. Hence, in our setting, the performance showed significant improvement only in the 1D dependency attention capturing module.

Model	I3D
With 1D dependency attention capturing module	0.7476
With affinity matrix module	0.6937
With increasing complexity module	0.7283

Table 5.1.2 AUC score of our model with different attention module

The performance of all models was also illustrated with the ROC curve as shown in figure 5.1.1. From figure 5.1.1, we found out that the despite our model with 1D dependency attention capturing module yielded the best result, the curve was similar to all other models. However, we observed that the true positive rate of our model with 1D dependency attention capturing module was higher than all other model. This could be the factor that caused its AUC score to be the highest.

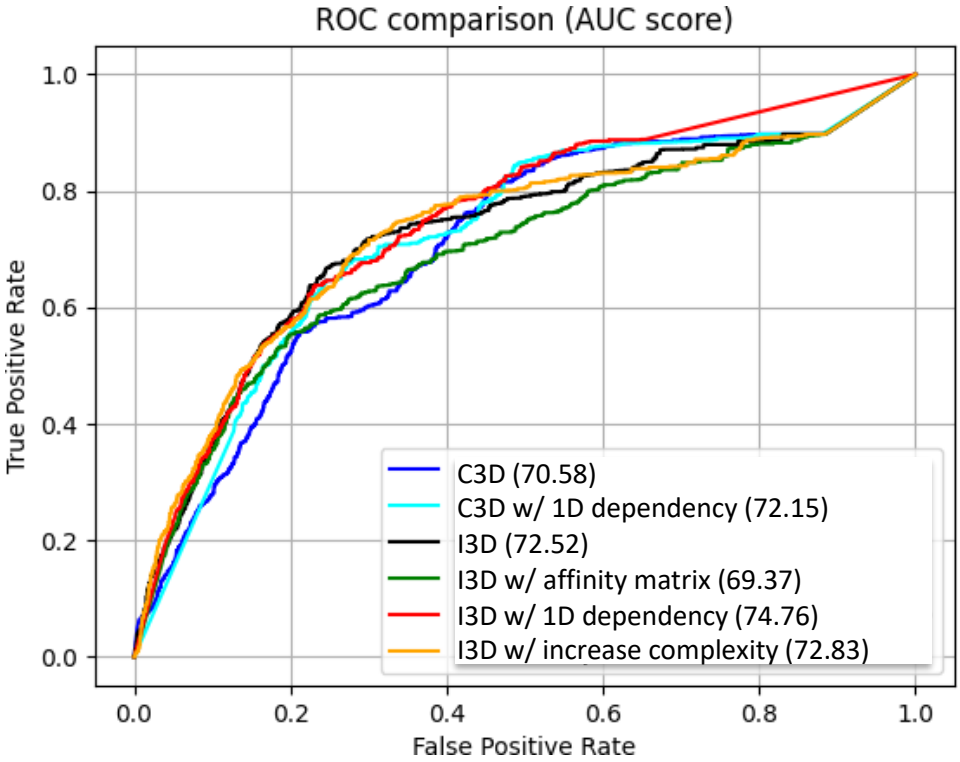


Figure 5.1.1 ROC Curve

We also plotted the precision-recall curve for each anomaly classes to compare the performance of our model with the baseline model as shown in figure 5.1.2 and figure 5.1.3. Based on the graph in figure 5.1.3, we found out that our model with 1D dependency attention

capturing module had overall higher precision and recall than the baseline model. Baseline model was the most precise in the temporal segment of stealing class whereas our model with spatiotemporal was the most precise in predicting the temporal segment of assault class. For recall, baseline model had the highest recall for fighting class and our model with 1D dependency attention capturing module had the highest recall for assault class. Baseline model had lowest precision in abuse class and lowest recall in vandalism class. Whereas our model with 1D dependency attention capturing module had lowest precision in explosion class and lowest recall in shoplifting class.

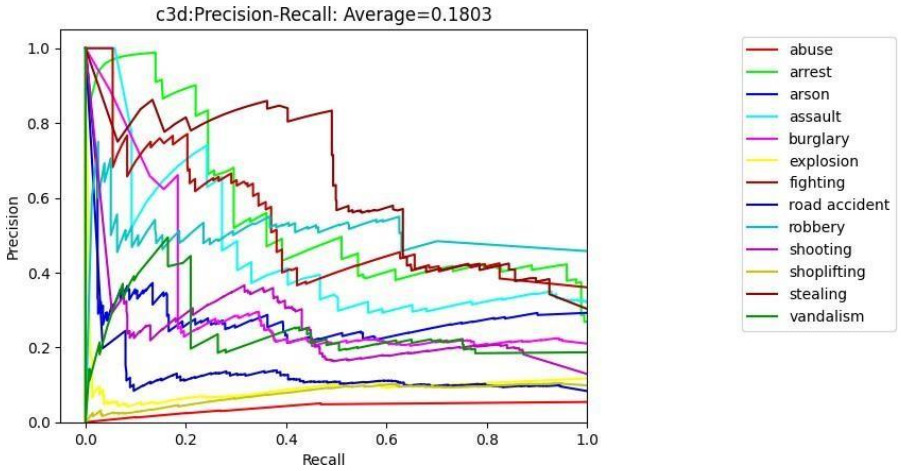


Figure 5.1.2 Precision recall curve of baseline model

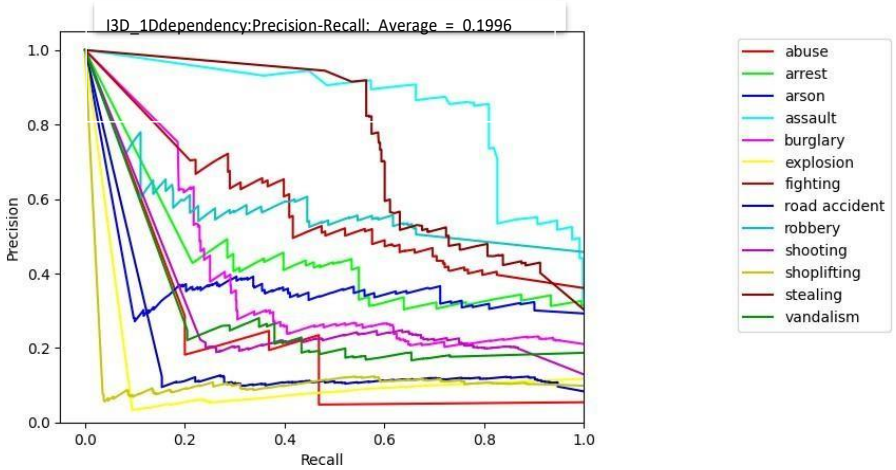


Figure 5.1.3 Precision recall curve of our model with spatiotemporal attention module

Figure 5.1.4 showed an example of misclassified temporal segment as burglary scene by baseline model. The reason for this misclassification was because the scene was too dark (night setting). However, our model successfully predicted high anomaly score in the

temporal annotation region as shown in figure 5.1.5. This showed that including 1D dependency attention capturing module could overcome the darkness limitation of baseline model.

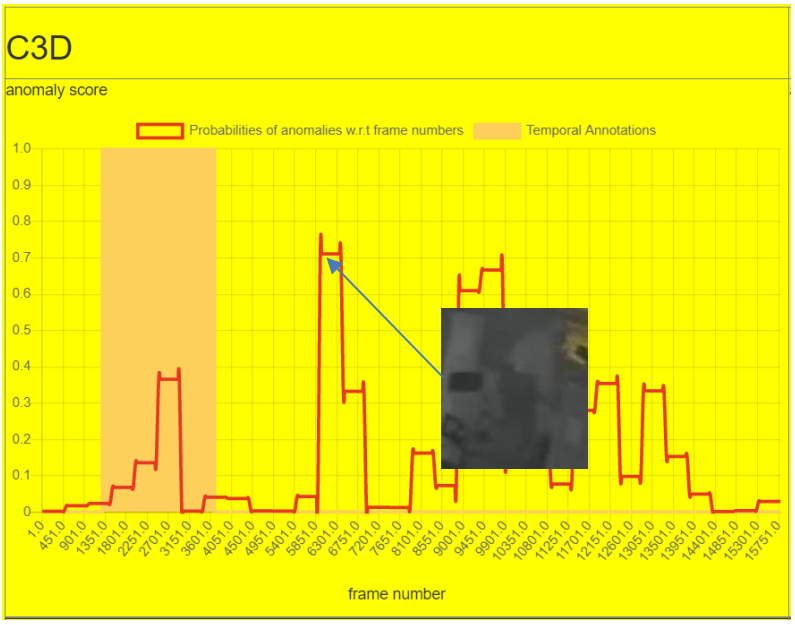


Figure 5.1.4 Failure case of baseline model

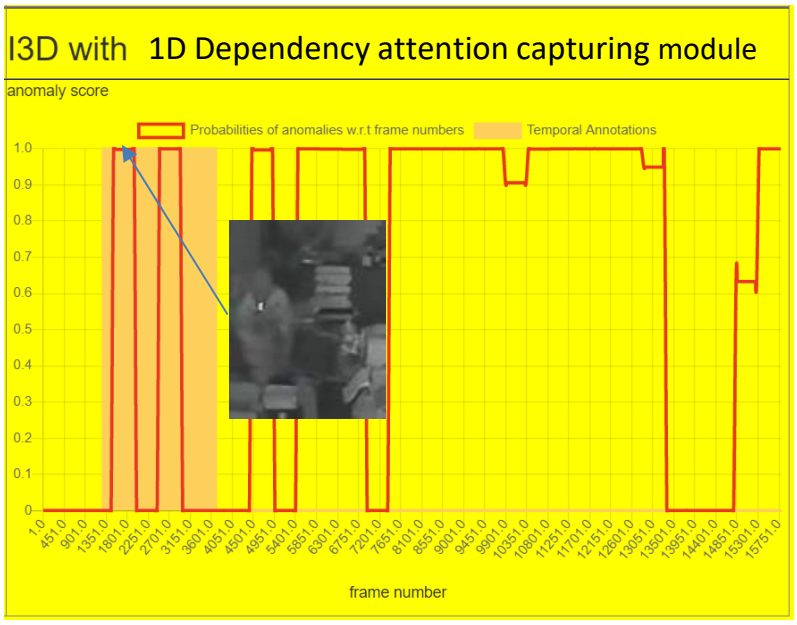


Figure 5.1.5 Success case of our model

Figure 5.1.6 and figure 5.1.7 showed the comparison between prediction without 1D dependency attention capturing module and with spatiotemporal attention module. From the plot in both figure 5.1.6 and figure 5.1.7, we observed that the curve of the one without 1D

dependency attention capturing module was smoother than the one with spatiotemporal attention module. This was because the 1D dependency attention capturing module assigned heavy weight to the anomaly segments. Hence the score gap between normal segment and anomaly segment was larger.

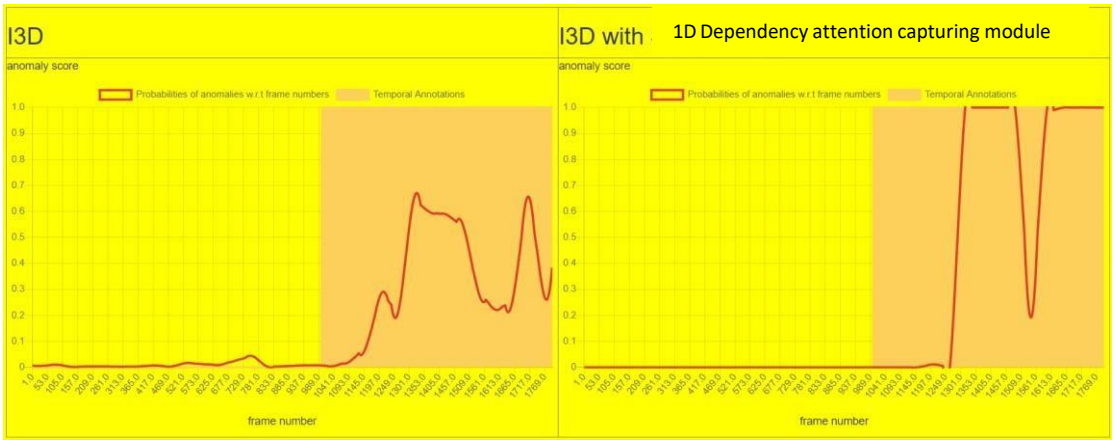


Figure 5.1.6 Sample prediction in arson class

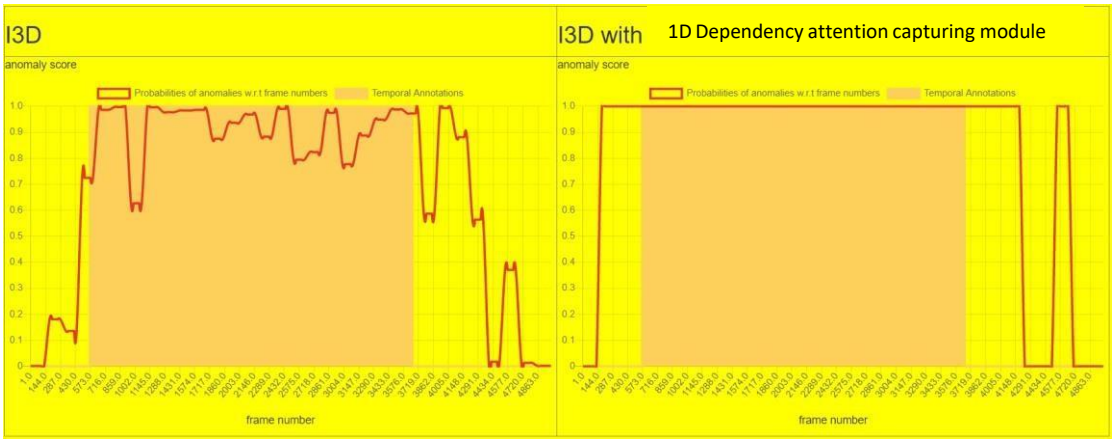


Figure 5.1.7 Sample prediction in stealing class

5.2 Error analysis

The following part of the report discussed the error analysis of our model with spatiotemporal attention module. We visualized the score predicted for each temporal segment on a graph. The model was expected to predict high anomaly score for anomaly temporal segment and low anomaly score for normal temporal segment. The area highlighted in orange on the graph was the ground truth anomaly temporal segment which was expected to have high anomaly score.

5.2.1 Successful case

Figure 5.2.1 – figure 5.1.11 were the examples of correctly predicted anomaly segment. Our model successfully predicted high anomaly score for the anomaly segment and low score for normal segment in all these examples.

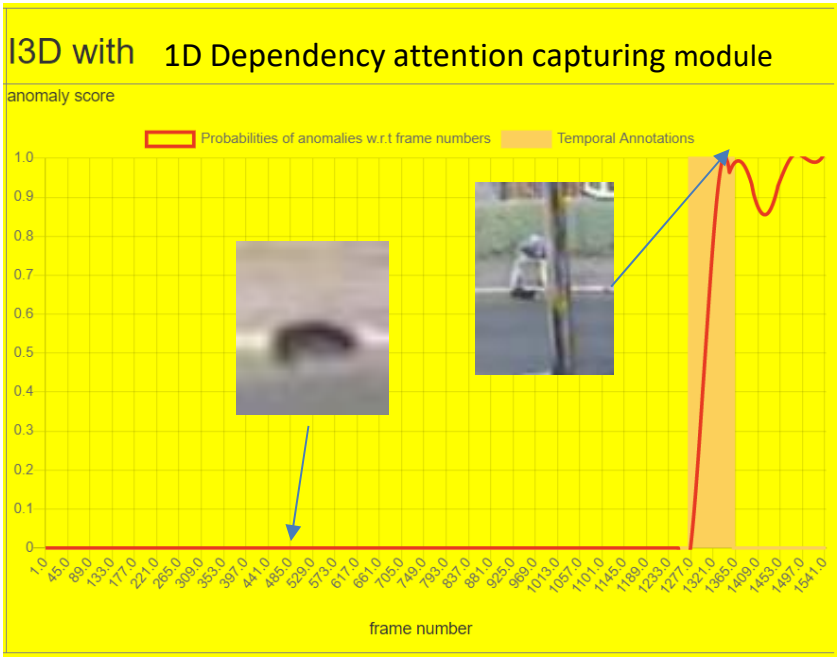


Figure 5.2.1 Success case in abuse class

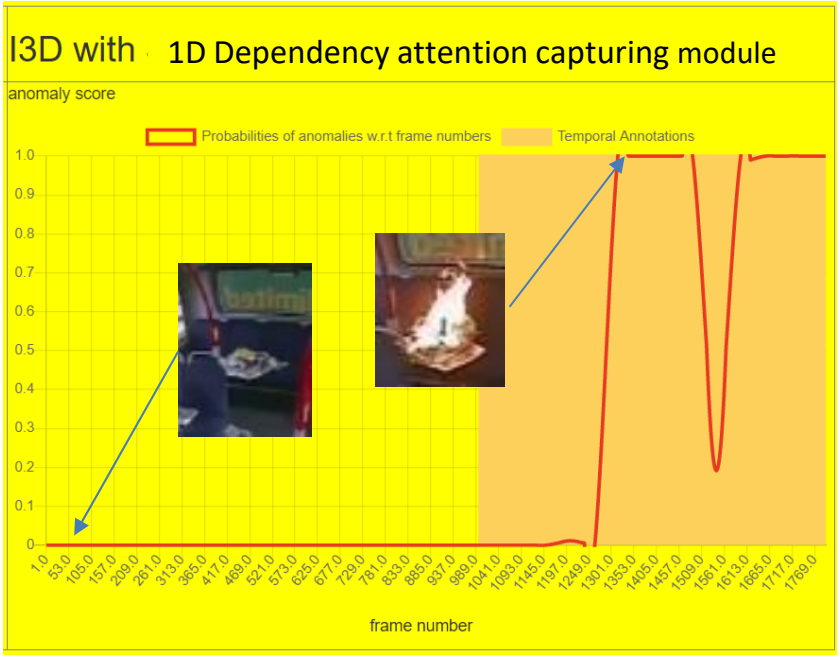


Figure 5.2.2 Success case in arson class

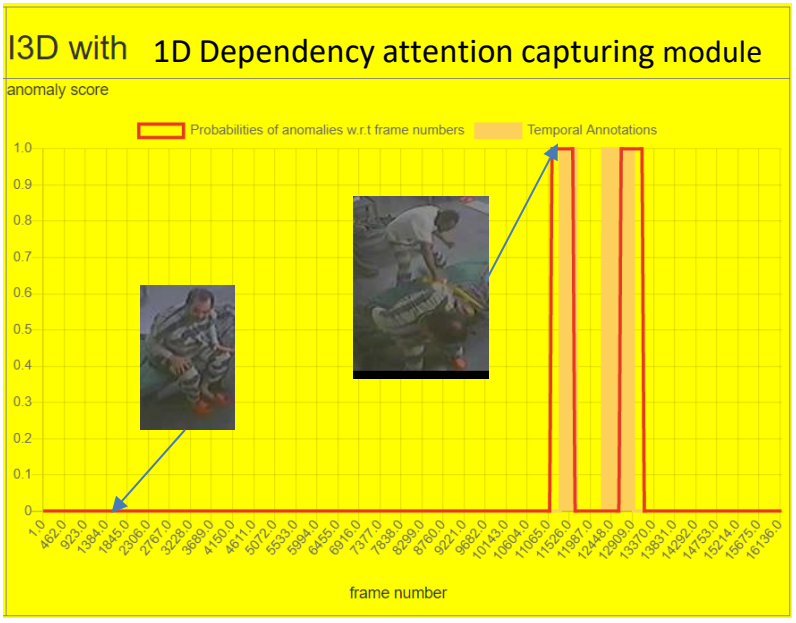


Figure 5.2.3 Success case in assault class

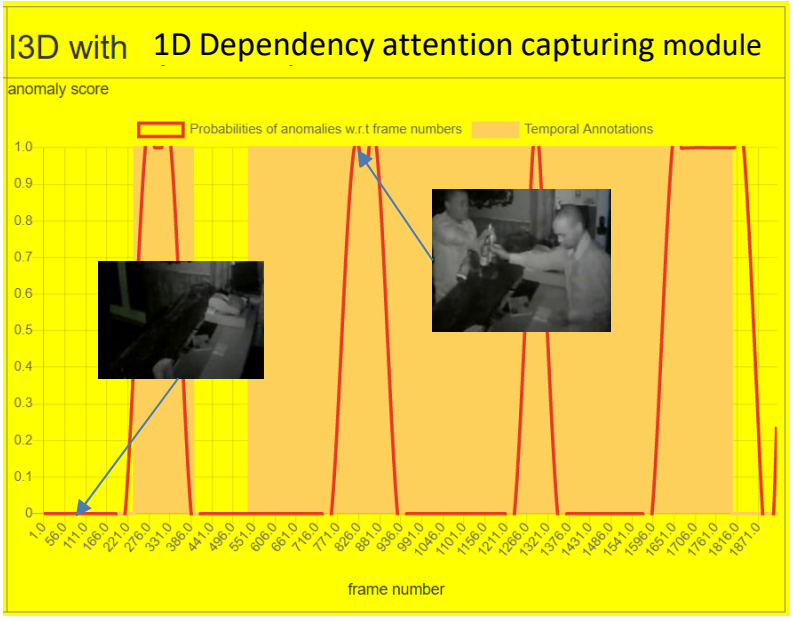


Figure 5.2.4 Success case in burglary class

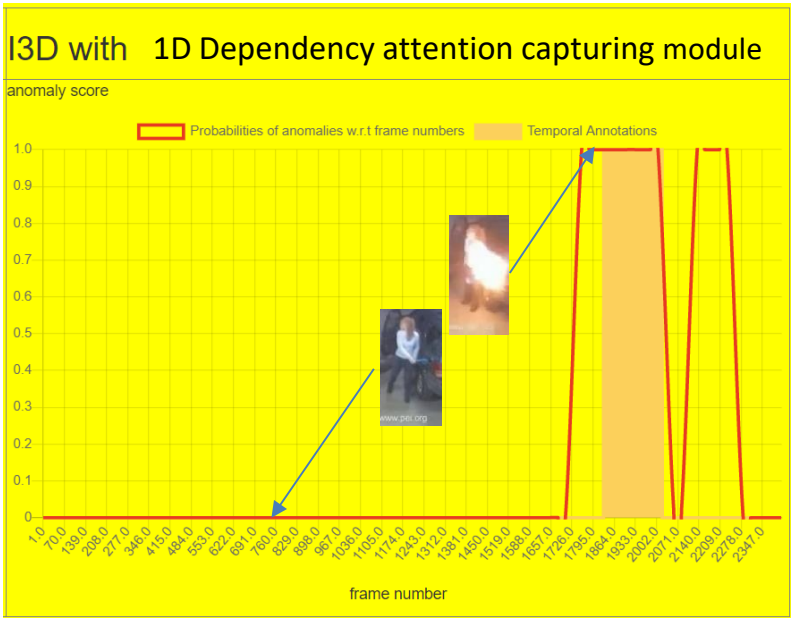


Figure 5.2.5 Success case in explosion class

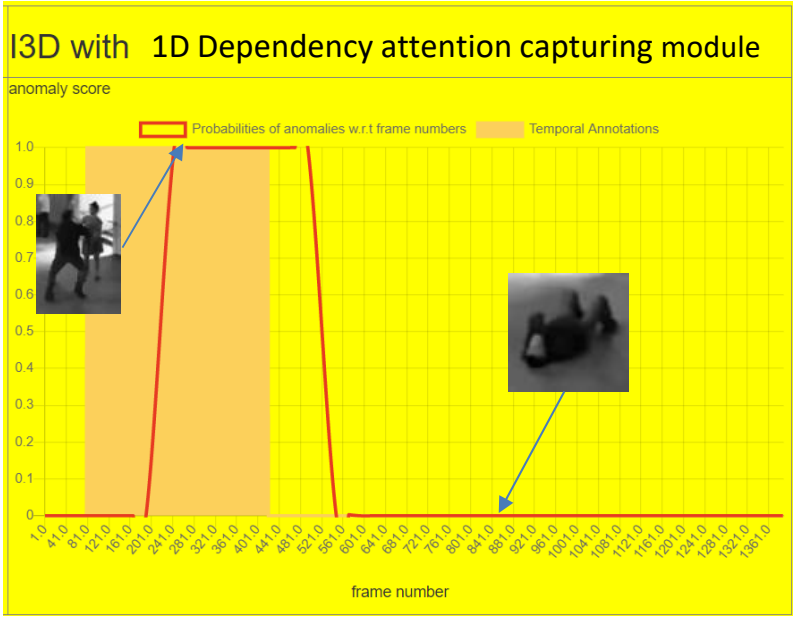


Figure 5.2.6 Success case in fighting class

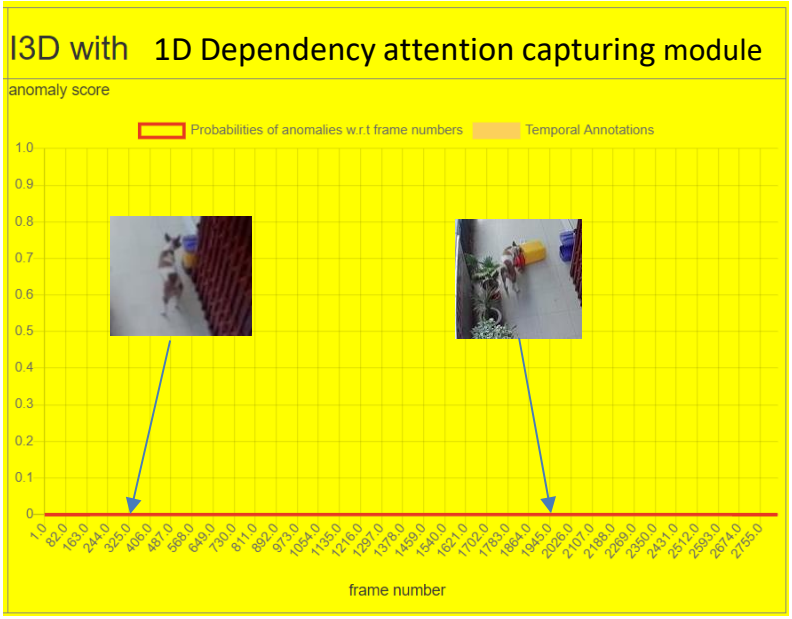


Figure 5.2.7 Success case in normal class

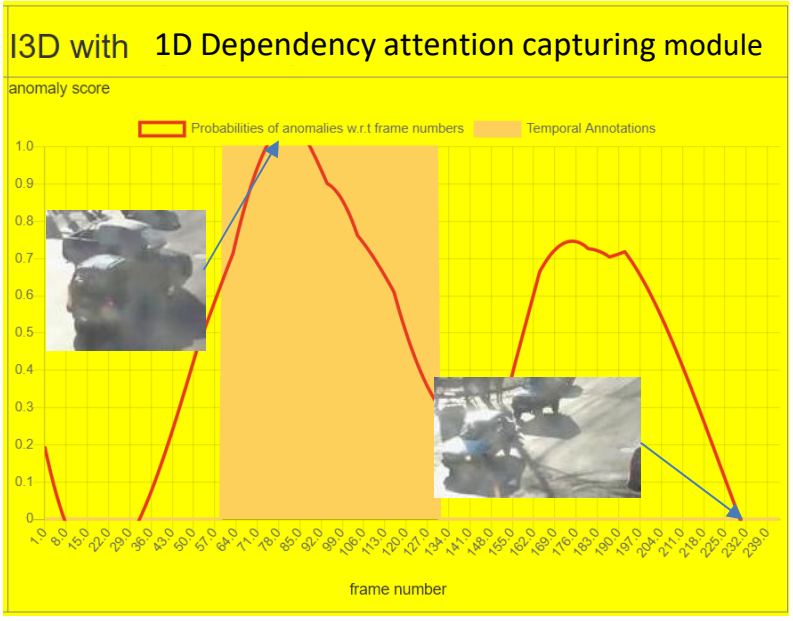


Figure 5.2.8 Success case in road accident class

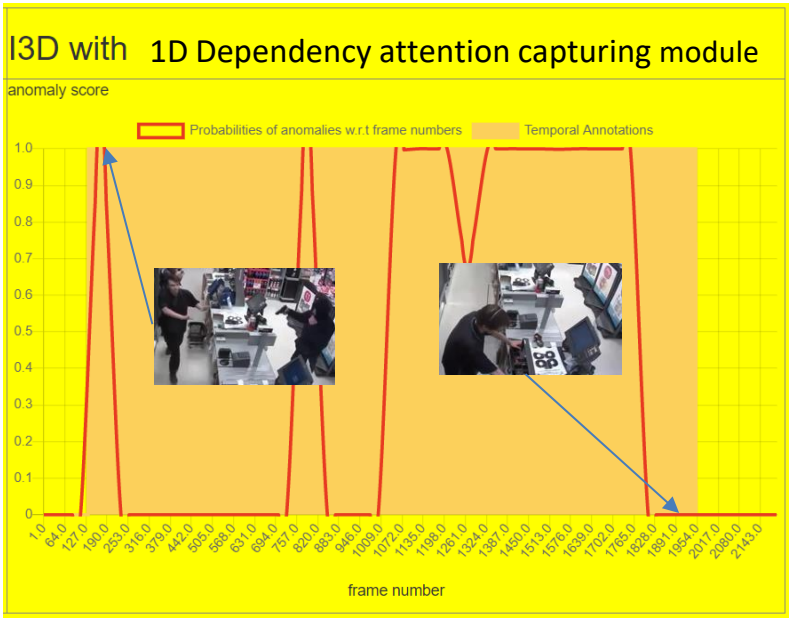


Figure 5.2.9 Success case in robbery class

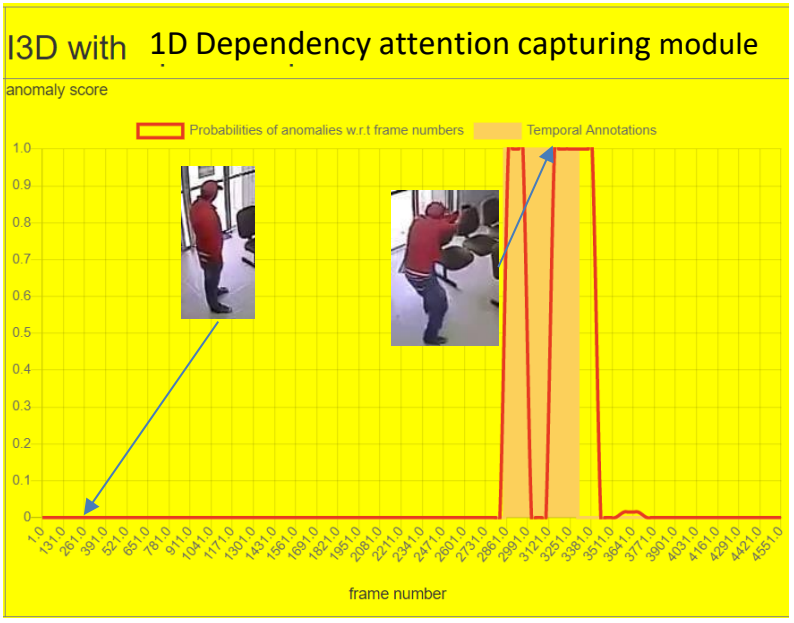


Figure 5.2.10 Success case in stealing class

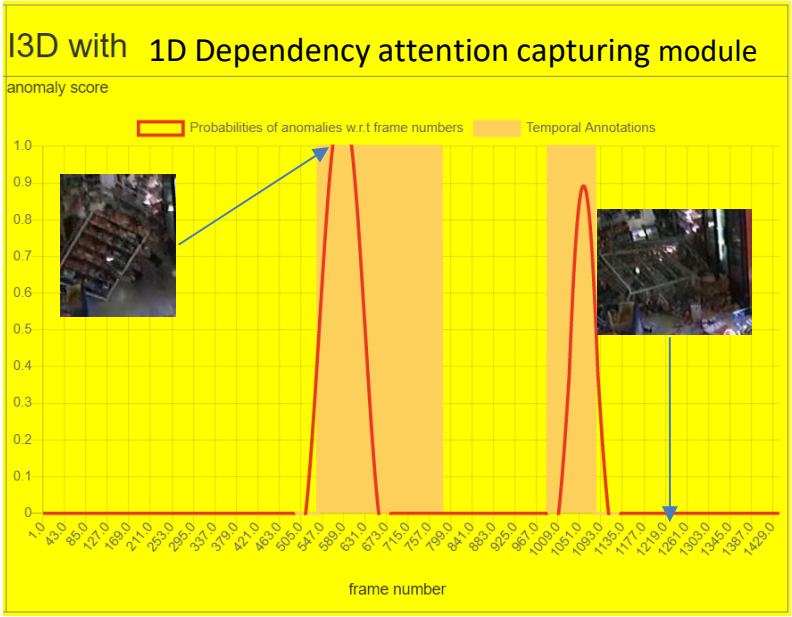


Figure 5.2.11 Success case in vandalism class

Based on figure 5.2.12, we observed that our model predicted high anomaly score before the temporal annotation. Despite, our model still able to predict high anomaly score in the temporal annotation too, so this was also considered as success case. From the sample frames on figure 5.2.12, we found out that our model predicted high anomaly score when the cars were near to each other which was what happened during temporal annotation. Hence, the model mistreated the first part of the temporal segment as arrest action too.

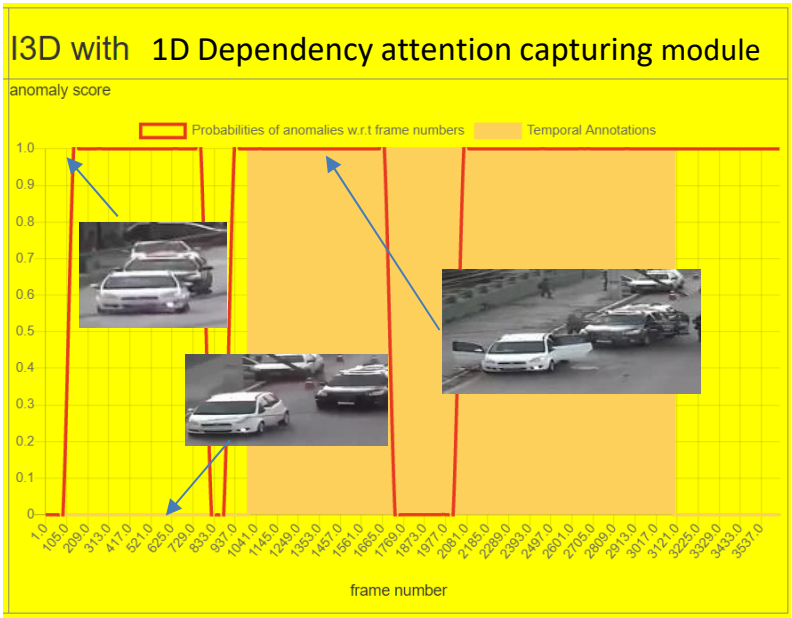


Figure 5.2.12 Success case of arrest class

5.2.2 Failure case

Figure 5.2.13 to figure 5.2.18 were examples of incorrectly predicted anomaly segment. Our model either predicted high anomaly score at the wrong temporal segment or low anomaly score throughout the whole video as if the crime scene was normal video.

Based on figure 5.2.13, we observed that the crime action occurred in not so obvious region in the frame, so our model failed to predict high anomaly score during the temporal annotation. From the sample frame shown on figure 5.2.13, the arrest action during the temporal annotation appeared in a darker region in addition the police were wearing black uniform in the frame. Hence, the model failed to predict high anomaly score for that segment, because it was not obvious. Same goes to figure 5.2.14, the shooting action appeared in very small part of the frame, hence our model failed to predict high anomaly score within the temporal annotation.

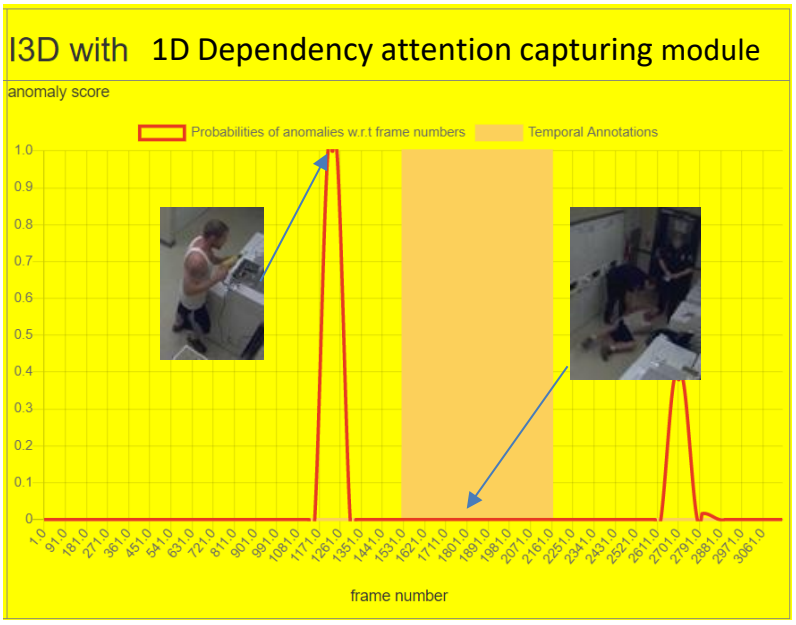


Figure 5.2.13 Failure case of arrest class

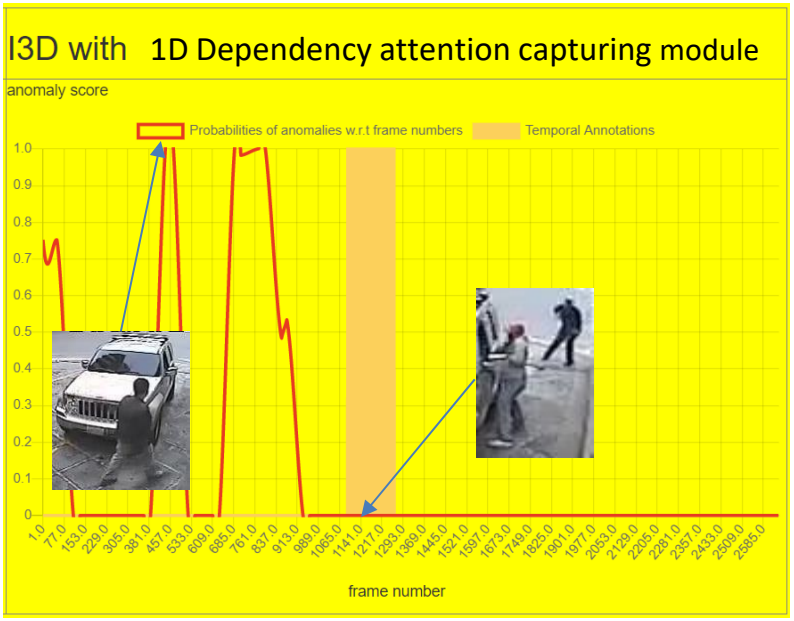


Figure 5.2.14 Failure case of shooting class

Based on figure 5.2.15 and figure 5.2.16, we observed that our model predicted low anomaly score throughout whole video. Our model treated the videos as they were normal videos. This was because the crime action was not captured by the surveillance video. Based on figure 5.2.15, the burglary action was blocked by the burglar in the surveillance video. Based on figure 5.2.16, the robbery action was blocked by the vehicle in the surveillance video.

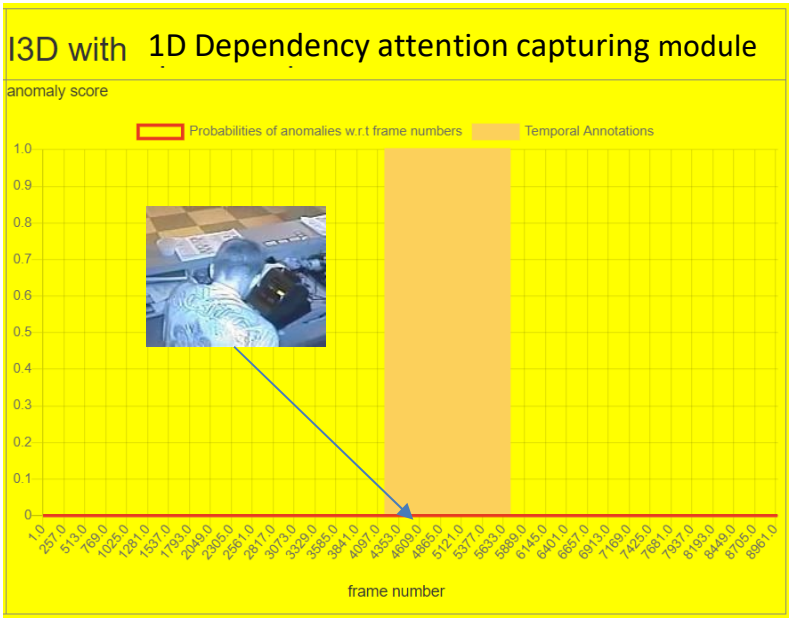


Figure 5.2.15 Failure case in burglary class

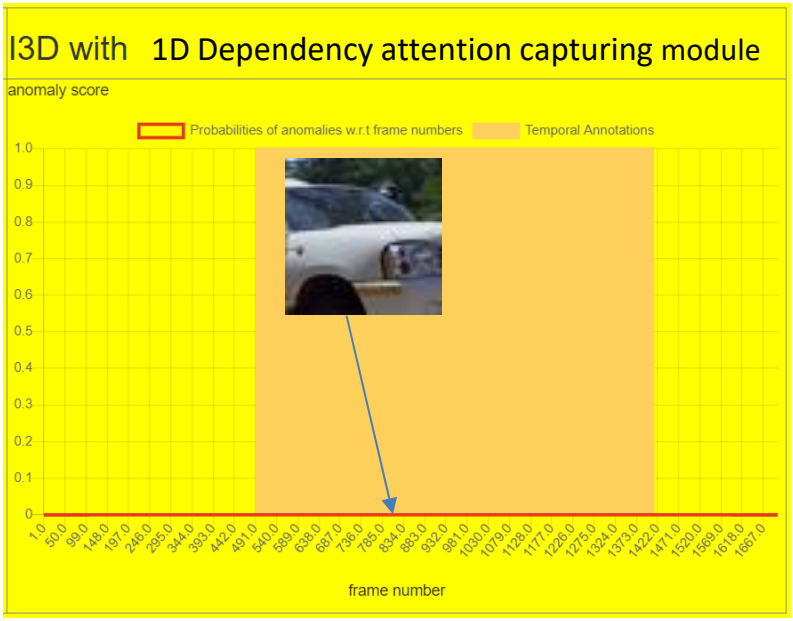


Figure 5.2.16 Failure case in robbery class

Figure 5.2.17 and figure 5.2.18 illustrated the examples of false alarm for the normal video. Based on the sample frame in figure 5.2.17, our model predicted high anomaly score because there was close interaction between the cashier and the customer. On the other hand, based on the sample frames in figure 5.2.18, our model predicted high anomaly score when there was huge movement in the frame such as van driving in or human cycling. Initially both of these videos were more like a still image without any action. Hence, when there was action, our model mistreated it as anomaly segment and predict high anomaly score for the segment.

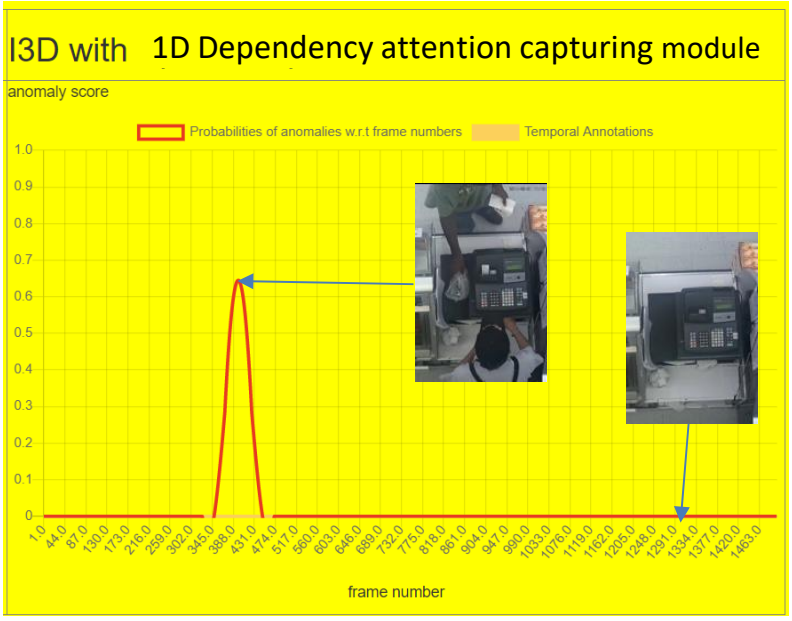


Figure 5.2.17 Failure case in normal class

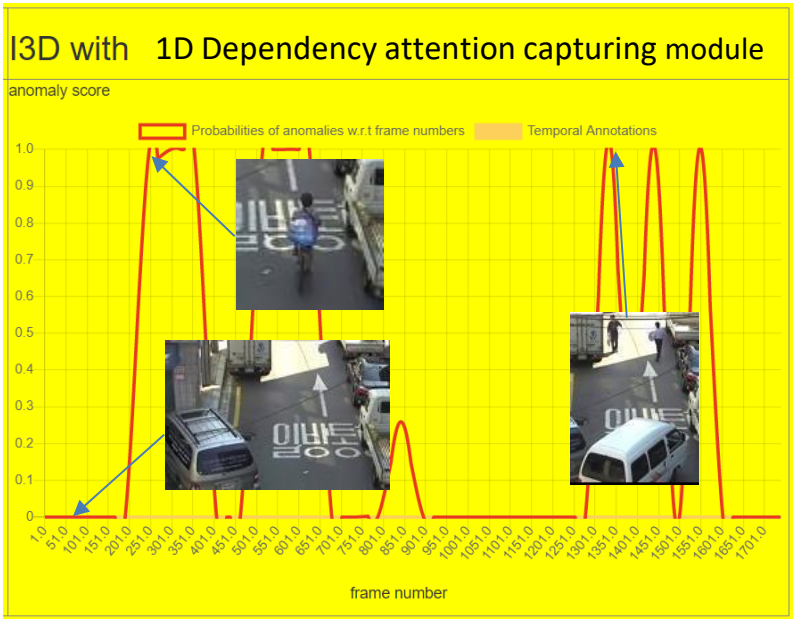


Figure 5.2.18 Failure case in normal class

CHAPTER 6

Conclusion

Based on the reimplementation of the MIL model proposed by [6], it showed that weakly annotated video dataset and training on both anomalous and normal video yielded a good performing crime scene detection system. Hence, in our project we resorted to semi supervised setting and also exploited both the normal and anomalous UCF crime dataset. However, there were still room for improvement for the MIL model proposed by [6].

To enhance the performance, we hypothesized that extracting 3D feature with deeper network would increase the performance of the network in our project. Besides, we also hypothesized that providing the MIL model with more descriptive and refined feature to learn on could improve the performance. Based on the result from our experiment, we have successfully achieved our objective and proven our hypothesis were correct. Extracting feature with deeper network and adding attention module on top of the feature yielded higher AUC score. Our implementation successfully outperformed the baseline model proposed by [6].

For future improvement, data augmentation can be done on certain action class to make the data more balanced. The dataset that we were using was imbalanced as some action class have 150 sample sizes, some only have 50 sample sizes. Training the model with balanced data might be able to improve the accuracy of all action classes. Also, augmenting the data can further prevent overfitting issue, so that we could use even deeper feature extractor. Besides, the attention module that we used in our project was self-attention module. For future work, we can experiment on the using different type of attention module such as spatial attention module and spatiotemporal attention module.

References

[1] E. Cosgrove, "One billion surveillance cameras will be watching around the world in 2021, a new study says," *CNBC,* 06 Dec 2019.

[2] P. Bischoff, "Surveillance camera statistics: which cities have the most CCTV cameras?," comparitech, 2021.

[3] Y. Gao, H. Liu, X. Sun, C. Wang and Y. Liu, "Violence detection using orented violent flows," 2016.

[4] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury and L. S. Davis, "Learning temporal regularity in video sequences," 2016.

[5] C. Lu, J. Shi and J. Jia, "Abnormal Event Detection at 150 fps in Matlab," 2013.

[6] W. Sultani, C. Chen and M. Shah, "Real-world Anomaly Detection in Surveillance Videos," 2018.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," IEEE, 2015.

[8] H. Wang and C. Schmid, "Action recognition with improved trajectories," 2013.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," 2014.

[10] A. Zisserman and C. Joao, "Quo Vadis, Action Recognition? A New Model and the," in *Cornell University*, 2018.

[11] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A Closer Look at Spatiotemporal COnvolutions for Action Recognition," 2018.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," 2014.

[13] Z. Qiu, T. Yao and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual network," 2017.

[14] A. Datta, M. Shah and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," in *IEEE*, Quebec City, QC, Canada, 2002.

[15] J. Kooji, M. Liem, J. Krijnders, T. Andringa and D. Gavrila, "Multi-modal human aggression detection," in *Computer Vision and Image Understanding*, Netherlands, 2016.

[16] S. Wu, B. Moore and M. Shah, "Chaotic invariants of langrangian particle trajectories for anomaly detection in crowded scenes," in *CVPR*, 2010.

[17] A. Bahsarat, A. Gritai and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *CVPR*, 2018.

[18] T. Hospedales, S. Gong and T. Xiang, "A markov clustering topic model for mining behaviour in video," in *ICCV*, 2009.

[19] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *CVPR*, 2009.

[20] I. Saleemi, K. Shafique and M. Shah, "Probabilistic modelling of scene dynamics for applications in visual surveillance," *TPAMI,* vol. 8, no. 31, pp. 1472-1485, 2009.

[21] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," 2009.

[22] X. Cui, Q. Liu, M. Gao and D. Metaxas, "Abnormal detection using interaction energy potentials," in *CVPR*, 2011.

[23] Y. Zhu, M. Nayak and A. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," in *IEEE Journal of Selected Topics in Signal Processing*, 2013.

[24] W. Li, V. Mahadevan and N. Vasoncelos, "Anomaly detection and localization in crowded scenes," in *TPAMI*, 2014.

[25] B. Zhao, L. Fei-Fei and E. Xing, "Online detection of unusual events in videos via dynamic sparse coding," *CVPR,* pp. 3313-3320, 2011.

[26] D. Xu, E. Ricci, Y. Yan and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *BMVC*, 2015.

[27] S. Lin, H. Yang, X. Tang, T. Shi and L. Chen, "Social MIL: Interaction-Aware for Crowd Anomaly Detection," Taipei, 2019.

[28] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic and C. Scmid, "ViViT: A Video Vision Transformer," in *ICCV*, 2021.

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	**Study week no.: 3**
Student Name & ID: Toh Yue Xiang & 18ACB01082	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Collaborative Batch Learning for Crime Scene Detection	

1. WORK DONE - Convert all video data into RGB frame data. - Feature extraction with I3D network. - Train MIL model with feature extracted from I3D feature extractor.
2. WORK TO BE DONE - Finetune the model to search for more suitable hyperparameter
3. PROBLEMS ENCOUNTERED - Much time taken to understand and developed the I3D feature extractor - Insufficient GPU memory which delayed the feature extraction.
4. SELF EVALUATION OF THE PROGRESS - Satisfy with current progress.

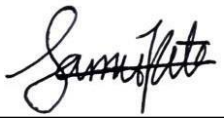
Supervisor's signature

Student's signature

Trimester, Year: Trimester 3, Year 3	**Study week no.: 6**
Student Name & ID: Toh Yue Xiang & 18ACB01082	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Collaborative Batch Learning for Crime Scene Detection	

1. WORK DONE - Performed hyperparameter tuning. - Research on attention module.
2. WORK TO BE DONE - Developed attention module. - Train the model with attention module.
3. PROBLEMS ENCOUNTERED - Selecting the suitable attention module to be implemented.
4. SELF EVALUATION OF THE PROGRESS - Satisfy with current progress.

Supervisor's signature



Student's signature

Trimester, Year: Trimester 3, Year 3	**Study week no.: 9**
Student Name & ID: Toh Yue Xiang & 18ACB01082	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Collaborative Batch Learning for Crime Scene Detection	

1. WORK DONE

- Developed multiheaded self-attention module.
-Trained MIL model with I3D feature with weight distribution assigned by self-attention module developed.

2. WORK TO BE DONE

- Hyperparameter tuning.
- Develop web app.

3. PROBLEMS ENCOUNTERED

- More GPU memory was needed after multiheaded self-attention module was included.
- The training time took much longer.

4. SELF EVALUATION OF THE PROGRESS

- Practice more on coding.
- Satisfied with current progress.

Supervisor's signature

Student's signature

Trimester, Year: Trimester 3, Year 3	**Study week no.: 12**
Student Name & ID: Toh Yue Xiang & 18ACB01082	
Supervisor: Ts Dr Tan Hung Khoon	
Project Title: Collaborative Batch Learning for Crime Scene Detection	

1. WORK DONE
- Hyperparameter tuning.
- Developed web app user interface design.

2. WORK TO BE DONE
- Developed functioning web app for prediction.
- Error analysis.
- Report writing.

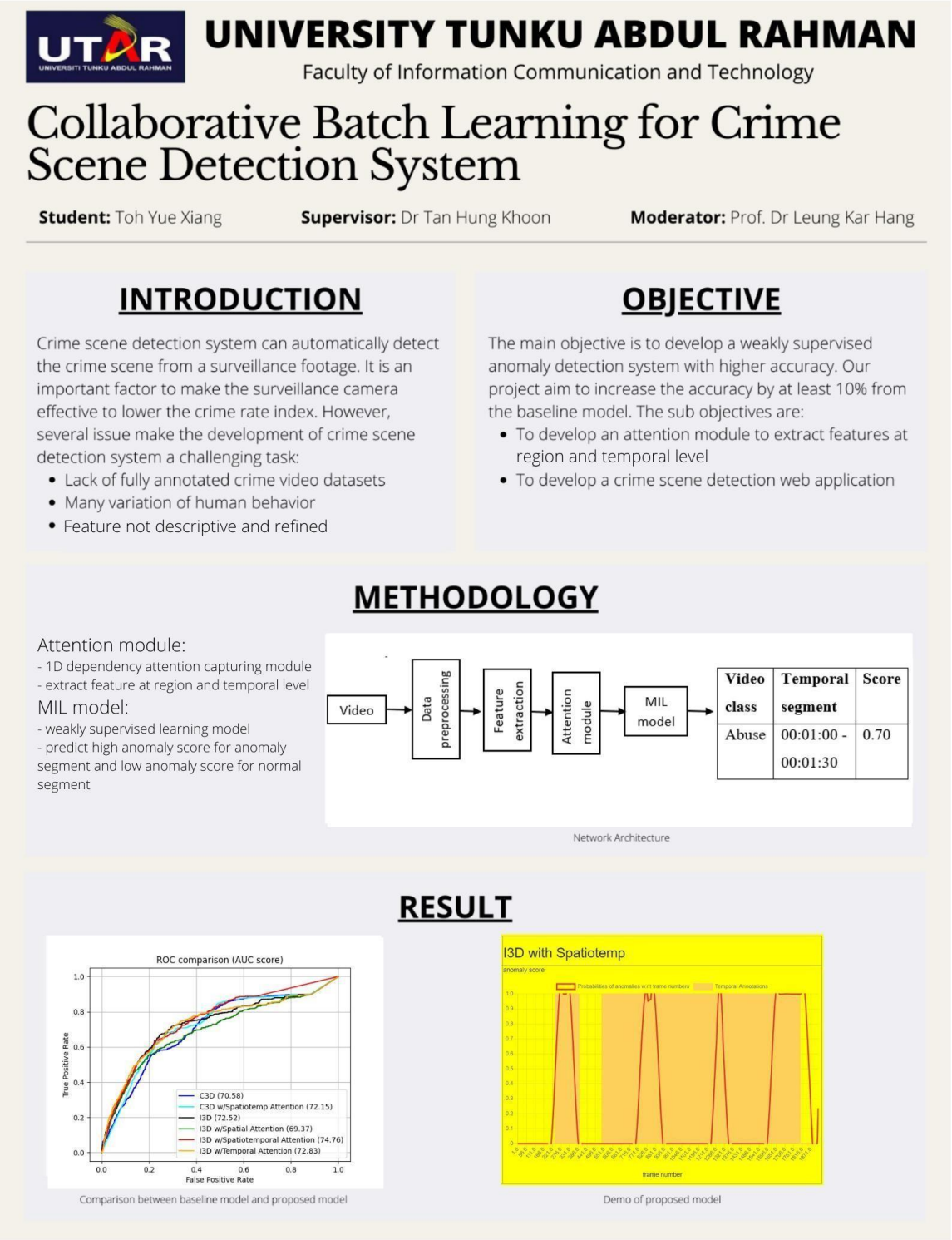
3. PROBLEMS ENCOUNTERED
- The performance initially did not improve much from baseline model.
- The implementation had errors.
- Took time to restudy front end web development.

4. SELF EVALUATION OF THE PROGRESS
- Need to be more careful when coding.
- The progress is acceptable but can be better.

Supervisor's signature

Student's signature

POSTER



UNIVERSITY TUNKU ABDUL RAHMAN

Faculty of Information Communication and Technology

Collaborative Batch Learning for Crime Scene Detection System

Student: Toh Yue Xiang **Supervisor:** Dr Tan Hung Khoon **Moderator:** Prof. Dr Leung Kar Hang

INTRODUCTION

Crime scene detection system can automatically detect the crime scene from a surveillance footage. It is an important factor to make the surveillance camera effective to lower the crime rate index. However, several issue make the development of crime scene detection system a challenging task:

- Lack of fully annotated crime video datasets
- Many variation of human behavior
- Feature not descriptive and refined

OBJECTIVE

The main objective is to develop a weakly supervised anomaly detection system with higher accuracy. Our project aim to increase the accuracy by at least 10% from the baseline model. The sub objectives are:

- To develop an attention module to extract features at region and temporal level
- To develop a crime scene detection web application

METHODOLOGY

Attention module:
- 1D dependency attention capturing module
- extract feature at region and temporal level

MIL model:
- weakly supervised learning model
- predict high anomaly score for anomaly segment and low anomaly score for normal segment

Video → Data preprocessing → Feature extraction → Attention module → MIL model →

Video class	Temporal segment	Score
Abuse	00:01:00 - 00:01:30	0.70

Network Architecture

RESULT

ROC comparison (AUC score)

True Positive Rate

False Positive Rate

C3D (70.58)
C3D w/Spatiotemp Attention (72.15)
I3D (72.52)
I3D w/Spatial Attention (69.37)
I3D w/Spatiotemporal Attention (74.76)
I3D w/Temporal Attention (72.83)

Comparison between baseline model and proposed model

I3D with Spatiotemp

anomaly score

Probabilities of anomalies w.r.t frame numbers

Temporal Annotations

frame number

Demo of proposed model

PLAGIARISM CHECK RESULT

Document Viewer

Turnitin Originality Report

Processed on: 19-Apr-2022 15:50 +08
ID: 1814335017
Word Count: 6635
Submitted: 1

Collaborative Batch Learning for Crime Scene ... By Toh Yue Xiang

Similarity Index	Similarity by Source	
3%	Internet Sources:	2%
	Publications:	1%
	Student Papers:	0%

include quoted | include bibliography | excluding matches < 8 words | mode: quickview (classic) report | Change mode | print | download

1% match (Internet from 24-Jan-2022)
http://openaccess.hacettepe.edu.tr:8080

<1% match (publications)
Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, Lin Chen. "Social MIL: Interaction-Aware for Crowd Anomaly Detection", 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019

<1% match ()
Zheng, Yin-Dong, Liu, Zhaoyang, Lu, Tong, Wang, Limin. "Dynamic Sampling Networks for Efficient Action Recognition in Videos", 'Institute of Electrical and Electronics Engineers (IEEE)', 2020

<1% match (student papers from 14-Apr-2022)
Submitted to Universiti Tunku Abdul Rahman on 2022-04-14

<1% match (Internet from 12-Feb-2019)
https://www.nwmo.ca/~/media/Site/Reports/2017/12/05/10/53/Mapping_Hornepayne_APM_REP_01332_0207.ashx?la=en

<1% match (publications)
"Neural Information Processing", Springer Science and Business Media LLC, 2017

<1% match ()
Bialobrzeski, Robert Wetherill. "Optimization of a SEGS solar field for cost effective power output", Georgia Institute of Technology, 2007

<1% match (Internet from 21-Dec-2021)
https://research.chalmers.se/publication/524155/file/524155_Fulltext.pdf

<1% match (Internet from 08-Jan-2019)
https://pt.scribd.com/document/203920054/Intellig-MANET

<1% match (Internet from 21-Jul-2020)
https://web.wpi.edu/Pubs/E-project/Available/E-project-102208-170809/unrestricted/Group_31_Final.doc

<1% match (Internet from 14-May-2019)
https://www.irjet.net/archives/V5/i3/IRJET-V5I3819.pdf

<1% match (publications)
Cheng Cheng, Pin Lv, Bing Su. "Spatiotemporal Pyramid Pooling in 3D Convolutional Neural Networks for Action Recognition", 2018 25th IEEE International Conference on Image Processing (ICIP), 2018

<1% match (publications)
Xiangzu Han, Fei Lu, Jianqin Yin, Guohui Tian, Jun Liu. "Sign Language Recognition Based on R(2+1)D With Spatial–Temporal–Channel Attention", IEEE Transactions on Human-Machine Systems, 2022

<1% match (publications)
Yang Liu, Keze Wang, Lingbo Liu, Haoyuan Lan, Liang Lin. "TCGL: Temporal Contrastive Graph for Self-supervised Video Representation Learning", IEEE Transactions on Image Processing, 2022

<1% match (publications)
Yongmei Zhang, Qian Guo. "A Human Action Recognition Algorithm in Dynamic Scene of Emergency Rescue", 2021 IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET), 2021

<1% match ()
Noordin, Nurul Hazlina. "FPGA Implementation Of A Multihop Wavelength Division Multiplexing (WDM) Ring Router Algorithm", 2004

turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Yue Xiang Toh
Assignment title:	FYP 202201
Submission title:	Collaborative Batch Learning for Crime Scene Detection
File name:	Toh_Yue_Xiang_18ACB01082_FYP2_turnitin.docx
File size:	3.51M
Page count:	39
Word count:	6,635
Character count:	35,007
Submission date:	19-Apr-2022 03:48PM (UTC+0800)
Submission ID:	1814335017

CHAPTER 1

Introduction

Surveillance camera is treated as a security measure in many countries to reduce the crime rate index. According to research, there are about 770 million surveillance cameras in use now and this figure is expected to increase to 1 billion by the end of 2021 [1]. The number of people per surveillance camera in China increase from 1 camera for 4.1 people in 2010 to 1 camera for 3.37 people in 2019 [2].However, it is found that the crime rate index is not correlated to the number of surveillance camera exist. One of the reasons to this is because the lack of artificial intelligence (AI) technology adaptation in the surveillance camera. The most used AI technology in surveillance currently is face recognition system. It is helpful in identifying the criminal identity only which is for solving crime case, but not in the detail ying the crime event which is for crime prevention. Hence, an intelligent surveillance camera system with anomaly detection system is needed to reduce crime rate index.

Anomaly detection system is a system to detect crime scene segment in the surveillance camera. For example, we input the surveillance footage, then the system will output the start and end time where the crime scene lies within the footage and what crime class the temporal segment belongs to. However, developing anomaly detection system is a challenging task in computer vision due to the absence of large fully annotated crime video dataset. Also, it is hard to judge an event as anomaly or normal sometimes. Hence, many researches have been ongoing to develop good performing crime scene detection system.

Some of the earlier work developed a specific type of anomaly detection system such as violence detection system [3]. Some other work exploited on normal video because normal video were available in larger quantity. For instance, deep auto-encoder based approach by [4] and dictionary based approach by [5].A some recent work proposed to use Multiple Instance Learning (MIL) model to develop an anomaly detection system [6]. There work exploited weakly labelled normal and abnormal video dataset and achieve significant performance. Their proposed approach was set up in weakly supervised setting which meant that only video level annotated dataset was needed to train their network.

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.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	TOH YUE XIANG
ID Number(s)	18ACB01082
Programme / Course	Computer Science
Title of Final Year Project	Collaborative Batch Learning for Crime Scene Detection

Similarity	**Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)**
Overall similarity index: 3 % **Similarity by source** Internet Sources: 2 % Publications: 1 % Student Papers: 0 %	
Number of individual sources listed of more than 3% similarity: 0	
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** ***Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.***	

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 Khoon

Date: 19/04/2022

Signature of Co-Supervisor

Name: -

Date: -

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY
(KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	18ACB01082
Student Name	TOH YUE XIANG
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.
N/A	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)
N/A	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
N/A	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: 19 April 2022