ISSN: xxxxxx

Crime Detection in Surveillance Camera Video with Multiple Instance Learning Vision Transformers

M.Alif Al Hakim1 , Tengku Laras Malahayati2 , Rakha Abid Bangsawan3 , Laksmita Rahadianti4
1Faculty of Computer Science, University of Indonesia, Depok, 16424, email: malif.al@ui.ac.id
2Faculty of Computer Science, University of Indonesia, Depok, 16424, email: tengku.laras@ui.ac.id
3Faculty of Computer Science, University of Indonesia, Depok, 16424, email: rakha.abid@ui.ac.id
4Faculty of Computer Science, University of Indonesia, Depok, 16424, email: laksmita@cs.ui.ac.id
Corresponding Author: M.Alif Al Hakim

ABSTRACT — Surveillance cameras or CCTV have been widely used to monitor and prevent crime. However, its effectiveness is often limited by human error, where security officers may not always focus on monitoring the area displayed on the CCTV monitor. This negligence in monitoring creates a gap in detecting crime immediately. Therefore, a system is needed that can help detect criminal acts automatically and in real-time. This study focuses on increasing the efficiency of detecting criminal acts through video surveillance using Video Vision Transformers trained with Multiple Instance Learning techniques and the Sliding Window approach. This study attempts to overcome the limitations of manual CCTV surveillance that is prone to human error in Indonesia. The UCF Crime dataset containing short criminal videos with a duration of around 4 minutes is used in this study. In addition, the dataset is also supplemented with personal videos obtained by the research team.
The Video Vision Transformers model is used to extract features from the video dataset, and the predictor model is trained to predict the anomaly score or the level of strangeness of criminal acts in the video. The proposed system automatically detects criminal acts based on the anomaly score value of each frame. The proposed method successfully achieved an AUC (Area Under the Curve) score of 0.942 in the video level context and 0.749 in the frame level context, indicating its effectiveness in detecting criminal acts in CCTV video. This method is also compared with a feature extraction method using a Convolutional Networks-based model with two image streams. The results of this study are expected to provide a significant contribution to improving the quality of public services in the security sector by reducing human error, increasing supervision efficiency, and creating a safer environment from crime.

KEYWORDS — CCTV, Detection, Crime, Multiple Instance Learning, Sliding Window, Video Vision Transformers, Weakly Supervised

I. INTRODUCTION

A. Background of the Problem

A sense of security is one of the basic human needs. Security is defined by Maslow, as quoted by Potter and Perry, as a feeling of protection from external and internal threats or terror related to security. Security is a condition in which a person is free from physical and psychological injury and in a safe and peaceful condition [1]. Security can be obtained from various things, one of which is being free from crime.

Crime in Indonesia is an event that threatens public security and violates criminal law. There are several examples of crime cases that often occur in Indonesia, such as theft, kidnapping, mugging, fraud and murder. Crime data for 2024 (until the second week of May 2024) accessed from the Indonesian National Police shows five types of crimes, including aggravated theft of 18,447 cases, ordinary theft of 15,267 cases, assault of 15,164 cases, fraud/fraudulent acts of 13,931 cases, and drug crimes of 13,482 cases [2].

Central Statistics Agency, citing the Criminal Code (KUHP) of the Republic of Indonesia and In-

The International Classification of Crime for Statistical Purposes (ICCS), initiated by the United Nations Office on Drug and Crime (UNODC), creates a classification of crimes as shown in Table 1.

Judging from the time of the crime, the time span between 18:00-21:59 is the time when the highest number of crimes occur with 24,190 cases.
Crime also occurs frequently between 08:00-11:59 with a total of 23,338 cases [2].

The security situation and conditions in Indonesia are dynamic. The Central Statistics Agency of Indonesia in the 2023 Criminal Statistics Publication shows that in 2020 the total number of crime incidents in Indonesia reached 247,218 incidents. This figure decreased in 2021 with 239,481 incidents. The downward trend only lasted for a moment because in 2022 there were 372,965 crime incidents. The agency also stated that the crime rate showed the same pattern.

In 2020, the crime rate was 94 and in 2021 it decreased to 91. Unfortunately, the crime rate then increased to 137 in 2022. In the 2023 Crime Statistics Publication, information can also be obtained about the time interval for crimes (crime clock), where in 2020 it was 00.02'.07", then it became 00.02'.11" in 2021 and

ISSN: xxxxxx

getting shorter in 2022 to 00.01'.24".
The time interval for the occurrence of crimes is increasing
This shortening indicates an increase in the intensity of
criminal incidents [3].

The relatively high crime rate in Indonesia as described above makes security a necessity. and preventing crime becomes important, one of them is by using a surveillance camera device, or what is commonly known as Closed-Circuit Television (CCTV). Stutzer and Zehnder, as quoted by Kurnia BP said that CCTV devices play a role as a supervisor and scout who can monitor the situation and real time conditions which are also capable of suppressing criminal acts to occur [4]. Kurnia BP also stated that CCTV is also used by those on duty as surveillance officers where they monitor CCTV with maximum preventive measures to minimize and detect suspicious actions [4]. The device CCTV has a significant role in combating and prevent criminal acts from occurring and provide protection for the community in public and private spaces. The use of CCTV devices is considered capable of providing security guarantees for the community as a form of effort preventive in early detection of criminal acts and emergency situation.

Unfortunately, CCTV surveillance and monitoring is manual by security supervisor is still not enough effective as an early detection effort, especially to detect criminal acts. This is caused by human factors. errors that can occur during supervision and monitoring done manually. For example, the supervisory officer security may not always focus on monitoring areas that displayed on a monitor displaying video footage CCTV when tired. Negligence in monitoring CCTV manually creates a gap in detecting criminal action.

Observing this gap, the research team assessed that These manual efforts to detect criminal acts are not efficient. To address this problem, the research team proposed the development of a system framework that can detect criminal acts automatically from CCTV videos. In its development, the research team built a dataset in the form of a collection of CCTV video recordings of criminal acts in Indonesia obtained from YouTube. Then, The dataset is used to develop a anomaly detection based model with two approaches different. With these two approaches, the detection system is expected to be able to detect criminal acts from video cameras. supervisor based on the anomaly score value on each frame. This system can be combined with an alarm notifier that can give warning to security supervisor about the possibility of criminal action when the anomaly score value quite high. Thus, the security supervisor can immediately follow up and verify the existence

No.	Crime Classification	Type of Crime
1.	Crime against life	Murder
2.	Crimes against the body	- Serious assault - Minor abuse - Domestic violence
3.	Crimes against morality	- Rape - Molestation
4.	Crimes against people's freedom	- Kidnapping - Employing children in underage
5.	Crimes against property rights/goods with use violence	- Theft by means of eras - Violent theft using a firearm (senpi) - Violent theft using sharp weapons (sajam)
6.	Crimes against property rights/goods	- Theft - Aggravated theft - Vehicle theft motorized - Destruction/destruction goods - Intentional burning - Receiving
7.	Drug related crimes	Narcotics and psychotropics
8.	Crimes related to fraud, embezzlement and corruption	- Fraud/fraudulent acts - Embezzlement - Corruption
9.	Crimes against public order general	- Against order general

TABLE I: Classification of Crimes
Source: BPS, 2023

the crime. This detection system is expected to be able to detect criminal acts in real time so that in the future can be implemented on CCTV. System development This is also expected to reduce human error in detect crime and assist security surveillance officers in conducting more efficient supervision.

B. Research Questions

Based on the background that has been explained, the research questions that the research team will propose are as follows.

1) How effective is the Multiple Instance approach? Learning and Video Vision Transformers with Slid-ing Sindow in detecting criminal acts in CCTV video?
2) What are the main challenges in implementing the approach? Multiple Instance Learning and Video Vision Transform-ers with Sliding Sindow to detect actions criminals in CCTV video?
3) How does the Multiple Instance approach perform? Learning and Video Vision Transformers with Sliding

ISSN: xxxxxx

Sindow when compared to the feature extraction approach using a Convolutional Two-Stream Network based model?

4) How can this video surveillance crime detection system be integrated with existing security systems to help improve efficiency in detecting crime?

C. Research Objectives

The following are the objectives to be achieved from this research.

1) To find out how effective the Multiple Instance Learning and Video Vision Transformers approaches are with Sliding Sindow in detecting criminal acts in CCTV camera videos.

2) Identifying the main challenges in implementing the Multiple Instance Learning and Video Vision Transformers approaches with Sliding Sindow to detect criminal acts in CCTV video.

3) To find out how the performance of the Multiple Instance Learning and Video Vision Transformers approaches with Sliding Sindow is compared to the feature extraction approach using a Convolutional Two-Stream Network-based model.

4) Understand how this crime detection system can be integrated with existing security systems to help improve efficiency in detecting crime.

D. Benefits of Research

This research is expected to provide benefits to parties who need it, both theoretically and practically, including the following.

1) Theoretical

Benefits • As a blueprint for creating an efficient security system in public facilities and can be a basis for other researchers in conducting similar research • This research is
expected to increase readers' insight and knowledge regarding incident detection in videos.

2) Practical Benefits

• Assisting the government in efforts to improve public facility services by increasing security at public facilities • Helping to increase the sense of comfort and security of the community at public facilities.

E. Research Limitations

The research limitations of the methods presented in this paper are as follows

1) The types of criminal acts that can be detected are only torture, arrest, arson, assault, road accident, burglary, explosion, fighting, stealing, shooting, shoplifting, and vandalism

2) The research only focuses on creating a de-taxi.

II. LITERATURE REVIEW

A. Weakly Supervised and Multiple Instance Learning

Anomaly detection is the process of finding unusual data or objects in a data set. The goal is to identify events or observations that differ from the majority pattern of the data [5]. Anomaly detection is used in a variety of fields, such as detecting fraud, disease, and crime. Often, anomaly detection is performed without complete and accurate data labels. This is also often the case in video anomaly detection, where labeling each video frame can be very expensive.

Weakly Supervised Learning is a way of training machine learning models where the data label information is imperfect [6]. There are three types of Weakly Supervised Learning: Incomplete Supervision (incomplete labels), Inexact Supervision (incomplete labels), and Inaccurate Supervision (labels are not always correct) [6]. In developing video anomaly detection models, the Inexact Supervision approach can reduce expensive labeling costs.

One way to implement Inexact Supervision is with the Multiple Instance Learning technique. This technique uses labels from larger data groups (bags) to label smaller data (instances) [7]. In anomaly detection, a bag is considered to have an anomaly if there is an instance or frame in it that contains an anomaly.

This approach has proven effective in several applications of anomaly detection models [8].

B. I3D Two Stream

I3D (Inflated 3D ConvNet) Two Stream is an artificial neural network model intended for analyzing videos.
This model is the result of the development of the 2D ConvNet model which is expanded (inflated) into three dimensions to process video data that has temporal and spatial information more effectively. Compared to the 2D ConvNet model, this model has a more complex dimension and has more parameters, which means that the model performance will be much better than the 2D ConvNet model because there is more sophisticated RGB processing and optical flow data processing [9].

In analyzing a video, the I3D Two Stream model has several steps. The diagram (Figure 1) explains the process carried out by the I3D Two Stream model.

First, the model receives two types of inputs, namely RGB Stream (the model takes information related to color and spatial details from each frame) and Optical Flow Stream (the model takes information related to the results of analyzing the differences between consecutive frames to capture motion). These two types of inputs are then processed through a 3D convolution process that is inflated from a 2D convolution model. There are two steps in this section, namely first the 2D convolution filter is inflated

ISSN: xxxxxx

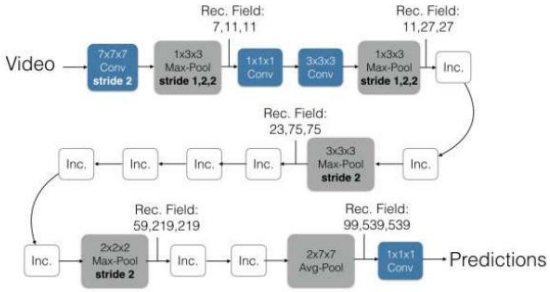

Figure 1: How the I3D Two Stream Model Works

into 3D by iteratively filtering it to
captures the movement patterns. Furthermore, the pooling layer also inflated to 3D to perform downsampling on
temporal and spatial dimensions. 3D pooling is performed to reduce data size and retain information
important in the video.

Next, each stream will be featured.
extraction. In the RGB stream, the model extracts spatial features of each frame. Meanwhile, in the Optical Flow stream,
The model extracts motion features from changes between frames.

After feature extraction from both streams,
The results will be combined for further analysis. The next step is to model using layers
fully connected to perform classification or detection
based on the combined features. For example, action/motion recognition detection, anomaly detection, etc.

C.Vision Transformers Video

Vision Transformers (VTs) is one of the branches
from the Transformers model architecture. This model was first introduced by Dosovitskiy et al. in the paper “An
Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” in 2020 [10]. The VTs model originates from the development of the Transformers model architecture, which
Initially often used in the field of language processing
natural language (NLP), for image recognition tasks. With Based on Transformers, VTs have a different approach
in capturing spatial relationships in images through
self-attention mechanism, which the approach lacks.
convolution used in Convolutional Neural Net-works (CNNs) [11].

In extending the use of VTs to video image recognition tasks, several adjustments have been made to
processing the temporal dimension in video content. The model such as VideoBERT [12] and TimeSformer [13] have shown promising results in understanding content.
video by combining spatial and temporal information.
VideoBERT adopts the BERT (Bidirectional En-coder Representations from Transformers) approach to process video frame sequence as text input to show
his ability to understand action and context in
videos without requiring annotations. Meanwhile, TimeSformer utilize self-attention spatially and temporally to process video frame sequences, resulting in better performance compared to CNN models.

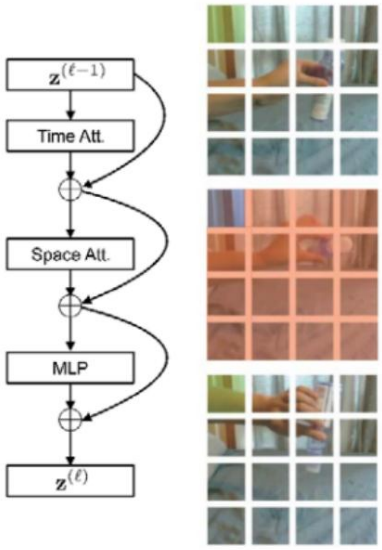

Figure 2: How Timesformer Works

III. METHODOLOGY

A. Solution Description

To overcome the problem of criminal acts in
community environment, the research team designed a crime detection system that can be integrated with surveillance cameras (CCTV). This system aims to
reduce the potential for negligence or human error that often occurs occurs in conventional CCTV surveillance. By utilizing the Multiple Instance Learning and Video
Vision Transformers, this system is able to analyze every video frames from surveillance cameras. When the level of abnormality or anomaly score detected is high enough
or exceeds a certain threshold, the system can provide alarm as an early warning. The threshold can be
adapted for each place to be monitored. As
for example, places that require tight security can
using a lower threshold to minimize false negatives. It is hoped that through the implementation of the system
With this detection, criminal acts can be identified more quickly. and handled by the authorities, thereby increasing
public safety and security.

B. Dataset

This study uses the UCF Crime dataset [14], which
contains 1900 videos with a total duration of 128 hours. This dataset covers 13 types of criminal acts: torture (abuse),
arrest, arson, assault, road accident, theft

(burglary), explosion, fighting, stealing, shooting, theft in

shoplifting, vandalism, and videos without permission

ISSN: xxxxxx

criminal acts. All videos with criminal acts are grouped into one anomaly group, while videos without criminal acts are grouped separately. In addition to the dataset, this study also utilizes an additional dataset collected by the research team itself. This additional dataset contains video recordings of surveillance cameras in Indonesia, with each type of criminal act represented by 2 videos. In this study, the UCF Crime dataset is used to train and test the model with video clip-level labels. Meanwhile, the additional dataset is used to evaluate the performance of the model in detecting anomalies in each frame so that labeling of each frame is done manually.

C. Research Methods

In general, the following is the flow of each method used in this study (Figure 3).

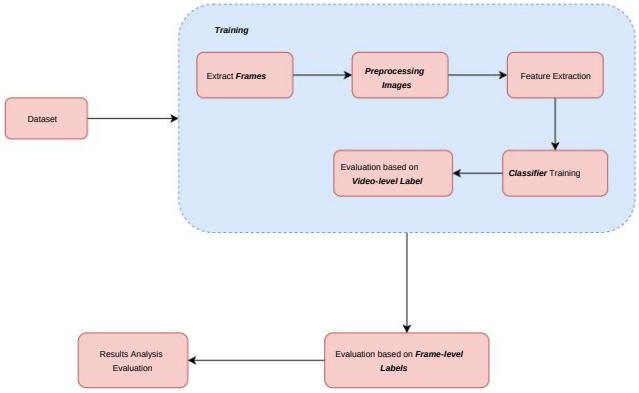

Figure 3: Research Flow

1) **Inflated 3D Convolution Networks Two Stream Method** The first step is to extract 30 frames per second from each video in the dataset. Each frame is then arranged so that its size is 320x240 pixels and an optical flow frame is produced that can capture the movement between frames. Next, features from the RGB stream and optical flow stream are extracted using the ResNet101 model that has been previously trained on the Kinetics-400 action recognition dataset [15]. After that, the extraction results are normalized.

The feature extraction results are then used to train the classifier model. The model used is an artificial neural network with 3 linear layers, ReLU activation function, and dropout on each layer except the last layer which uses the sigmoid function to produce probability values. The model is trained using the Multiple Instance Learning technique where video-level labels are used during training.

The model is then evaluated using video-level labels from the test portion of the UCF Crime dataset and the labels

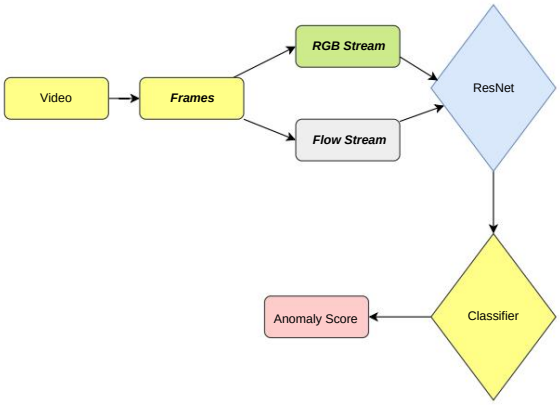

Figure 4: How the I3D Two Stream Method Works

frame level of additional datasets that the research team collected. In measuring model performance, the research team used the ROC AUC metric [16] to measure model performance. Furthermore, an analysis of model performance was carried out in terms of accuracy and inference time.

2) **Video Vision Transformers Method with Sliding Window**

Unlike previous approaches that used Convolutional Networks-based models, in this method the research team proposed a different approach based on the Transformers model. In this study, the TimeSformer model [13] was used, which is one of the Video Vision Transformers models [17] that has been trained on action recognition datasets, such as Kinetics-400 [15]. TimeSformer basically only has the ability to classify one video clip into a certain action category.

In this method, the model training process involves extracting sample frames from each video. Randomly, 8 frames are selected from each video. Then each frame is preprocessed, including changing the size to 224x224 pixels, scaling the pixel values to the range of 0 to 1, and normalizing the pixel values.

The processed frames are used to fine-tune the TimeSformer model and train a simple classifier model in the form of a single linear neuron with a sigmoid activation function. Model training is performed using the Multiple Instance Learning technique with video-level labels.

The model is then tested using video-level labels from the UCF Crime dataset test subset and frame-level labels from the supplementary dataset. To predict anomalies on each frame, a Sliding Window approach is used, where 8 frames are collected and used to

ISSN: xxxxxx

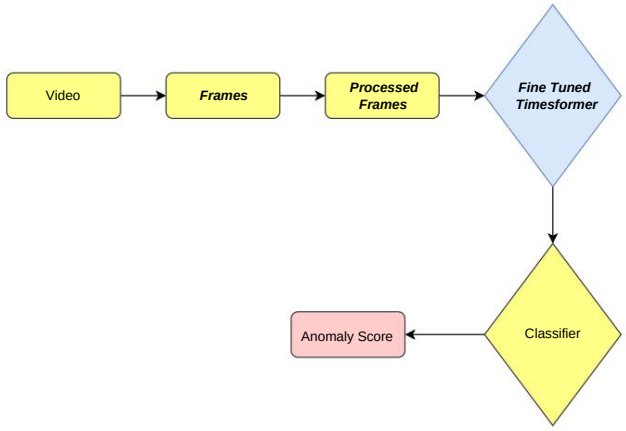

Figure 5: How the Vision Transformers Method Works

Anomaly score prediction. The next frame is predicted using the first to ninth frames,
and so on. Performance analysis is also performed on this method.

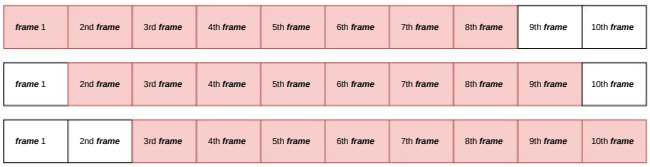

Figure 6: Sliding Window Visualization

D. Evaluation Metrics

This study uses the Area Under the Receiver metric Operating Characteristic Curve (AUC-ROC) [16] to measure the performance of the classification model at various thresholds. (threshold). The Receiver Operating Characteristic (ROC) is a probability curve that describes the relationship between true positive rate (TPR) and false positive rate (FPR). AUC is the area under the ROC curve, which indicates how well the model can distinguish between the positive and negative classes. negative. The higher the AUC value, the better the ability. model in distinguishing the two classes. This metric then used to measure the accuracy of each prediction method of detecting anomalies in each frame (frame-level) and the presence or absence of anomalies in a video (video-level).

IV. EXPERIMENTAL AND TESTING RESULTS

In this study, a convolutional network-based method trained using the Adam optimizer with a learning rate 0.001 for 75 epochs and batch size 30, and using the MultiStepLR scheduler. Meanwhile, the method based on Transformers trained for 3 epochs with Adam optimizer which has a learning rate of 0.00005 and a batch size of 2, and using a linear scheduler. In addition, in this study loss function used is binary cross entropy.

The following are the results of the research team's tests. get when both methods are tested on video-level labels and frame-level labels.

Method	AUC-ROC Score
I3D Two Stream Method	0.841
Video Vision Transformers Method	0.942

TABLE II: AUC ROC score video-level labels

Method	Average AUC-ROC Score
I3D Two Stream Method	0.515
Video Vision Transformers Method	0.749

TABLE III: AUC ROC score frame-level labels

It can be seen that the Video Vision Transformers method has better metric values compared to the I3D Two method Streams well on tests using video-level labels or frame-level labels.

In addition, the research team also conducted tests on inference speed of each method. For this purpose, it is carried out calculation of the average number of frames that can be processed per seconds by each method.

Method	**Frames** per second
I3D Two Stream Method	11.1
Video Vision Transformers Method	5.6

TABLE IV: AUC ROC Score frame-level labels

V. ANALYSIS OF EXPERIMENTAL AND TESTING RESULTS

The test results in Table II show that the method Video Vision Transformers (AUC-ROC score 0.942) outperforms the I3D Two Stream method (AUC-ROC score 0.841) in terms of prediction accuracy at the video level. This is indicates that the Video Vision Transformers model better able to distinguish between videos containing criminal acts and normal videos as a whole.

Next, in testing with frame-level labels (Table IV), the Video Vision Transformers method again shows better performance with an average AUC-ROC score of 0.749, compared to the I3D Two method. Stream which only reached 0.515. This advantage shows that Video Vision Transformers is more effective in identify criminal acts at the individual frame level, which means the model is more accurate in localizing specific moments when criminal acts occur in a video.

ISSN: xxxxxx

Overall, the results of this evaluation show the potential of the Video Vision Transformers method as a superior approach in developing a video-based crime detection system, especially in terms of accuracy. However, it should be noted that this method has a weakness in efficiency, as seen from the number of frames that can be processed per second which is less than the I3D Two Stream method.

In addition to differences in performance and efficiency, the I3D Two Stream method also has weaknesses in producing models with a high level of prediction confidence. This can be seen in Figure 8, which visualizes the anomaly score of one of the videos in the dataset. The video has

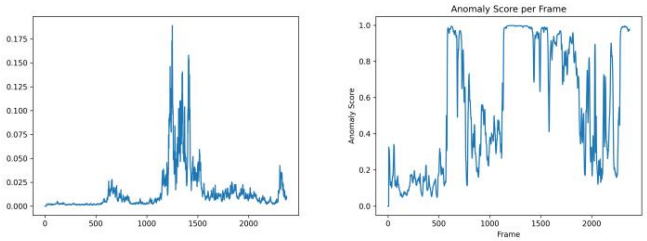

Figure 7: Anomaly Score of I3D Method (Left) and Transformers (Right)

two anomalous frame intervals, namely between frames 623 to 794 and frames 1185 to 1485. Although both models tend to detect the second anomalous event, the I3D Two Stream model fails to identify the first anomalous event. In addition, it is seen that the anomaly scores produced by the I3D Two Stream model are generally lower than those of the Video Vision Transformers model, even when an anomalous event occurs. Similar findings are also found in other videos in the dataset. This significant difference in anomaly scores indicates that the I3D Two Stream model is less able to clearly distinguish between normal frames and frames containing anomalies.

Figure 8: I3D Method Prediction Overview (Left) and Transformers (Right)

VI. CONCLUSION

This study shows that the Multiple Instance Learning and Video Vision Transformers with Sliding Window approaches are quite effective in detecting criminal acts. This approach also proves to be superior in terms of accuracy compared to convolutional network-based approaches, although there are still challenges in terms of frame processing efficiency. Therefore, it is important to have adequate computing resources when this technology is integrated in the field. In addition, in its application, it is important to adjust the anomaly score threshold based on the urgency of each monitored location to produce an optimal alarm notifier. In closing, it is hoped that this research can be useful for the development and management of security systems in the future.

REFERENCE

[1] AG Perry and PA Potter, “Fundamentals of nursing textbook vol, 2.” Egc, 2005.

[2] Republic of Indonesia National Police (POLRI). (2024) Crime data. [Online]. Available: https://pusiknas.polri.go.id/datajahat [3] Central Statistics Agency, “Criminal statistics 2023. vol. 14,” https://www.bps.go.id/id/publication/2023/12/12/5edba2b0f e5429a0f232c736/statistik-kriminal-2023.html, 2023, online; accessed 13 May 2024.

[4] GRCK BP, “The role of closed-circuit television (CCTV) surveillance cameras in counter-terrorism,” Jurnal Lemhannas RI, vol. 9, no. 4, pp. 100– 116, 2021.

[5] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM computing surveys (CSUR), vol. 41, no. 3, pp. 1–58, 2009.

[6] Z.-H. Zhou, “A brief introduction to weakly supervised learning,” National science review, vol. 5, no. 1, pp. 44–53, 2018.

[7] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, “Multi-ple instance learning: A survey of problem characteristics and applications,” Pattern Recognition, vol. 77, pp. 329–353, 2018.

[8] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Shen, “Batchnorm-based weakly supervised video anomaly detection,” arXiv preprint arXiv:2311.15367, 2023.

[9] Z. Wang, “Anomaly detection in surveillance videos based on two-stream inflated 3d conv net and weakly supervised learning,” in CIBDA 2022; 3rd International Conference on Computer Information and Big Data Applications, 2022, pp. 1–5.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, 2020.

[11] J. Bi, Z. Zhu, and Q. Meng, “Transformers in computer vision,” in 2021 IEEE International conference on computer science, electronic information engineering and intelligent control technology (CEI). IEEE, 2021, pp. 178–188.

[12] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “Videobert: A joint model for video and language representation learning,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7464–7473.

[13] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?” in ICML, vol. 2, no. 3, 2021, p. 4.

[14] Waqas Sultani, Chen Chen, Mubarak Shah. (2018) Real-world anomaly detection in surveillance videos. [On line]. Available: https://www.crcv.ucf.edu/projects/real-world/ [15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijaya-narasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., “The kinetics human action video dataset,” arXiv preprint arXiv:1705.06950, 2017.

[16] A.P. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422 [17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Luciÿ c, and C. Schmid, “Vivit: A video vision transformer,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836–6846.