IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 September 2023, accepted 23 November 2023, date of publication 4 December 2023,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3339379

RESEARCH ARTICLE

KianNet: A Violence Detection Model Using an Attention-Based CNN-LSTM Structure

SOHEIL VOSTA, (Student Member, IEEE), AND KIN-CHOONG YOW, (Senior Member, IEEE)
Department of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada

Corresponding author: Kin-Choong Yow (kin-choong.yow@uregina.ca)

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number DDG-2020-00034. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), numéro de référence DDG-2020-00034.

ABSTRACT Violent behaviour is always an important issue that threatens any society. Therefore, many organizations have used surveillance cameras to monitor such events to preserve public safety and mitigate potential harm. It is difficult for human operators to monitor the copious camera feed manually, however, automated systems are employed to enhance the accuracy of violence detection and reduce errors. In this paper, we propose a novel model named KianNet that effectively detects violent incidents inside recorded events by combining ResNet50 and ConvLSTM architectures with a multi-head self-attention layer. The utilization of ResNet50 enables robust feature extraction, while ConvLSTM makes it easier to take advantage of the temporal dependencies in the video sequences. Furthermore, the multi-head self-attention layer enhances the model's ability to focus on relevant spatiotemporal regions and their discriminatory capacity. Empirical investigations confirm that the proposed model outperforms its competitors by roughly 10 percent, achieving a 97.48% AUC on binary classification on the UCF-Crime dataset, and a 96.21% accuracy on the RWF dataset, surpassing Violence 4D.

INDEX TERMS Violence detection, anomaly detection, computer vision, ResNet, ConvLSTM, attention mechanism, multi-head self-attention, UCF-Crime, RWF, vision saccade.

I. INTRODUCTION

With the growing challenges in public safety and security, the demand for comprehensive public safety monitoring via video surveillance cameras has significantly increased.

However, the abundance of video data generated by these surveillance cameras, associated with the limited availability and diversity of anomalous events such as violence, theft, or other types of crimes, presents a notable challenge to detecting abnormal behaviours. Manual monitoring of this expansive data is impractical and labour-intensive and tends to cause errors due to human visual fatigue. Thus, this highlights the urgent requirement for effective and automated systems for detecting violence.

Like any technological advance, the applications of these systems can be manifold in different aspects. One of the most significant societal implications of automated violence detection systems is improving public safety and proactivity. In that, surveillance systems that can automatically detect signs of violence or aggressive behaviour have the potential to save lives. For instance, if a system can detect a potential act of violence in a public space, rapid response units can be deployed before situations escalate, thereby preventing harm. In contrast to traditional surveillance methods that rely on human monitoring, automated systems can continuously monitor numerous feeds simultaneously, leading to proactive interventions rather than waiting for something severe to happen and reacting.

One of the principal methodologies utilized in video classification is Supervised Learning, widely used in violence detection (VD) models for distinguishing violent behaviours from normal ones. This method can efficiently use labelled data and learn unique characteristics for each category. However, when it comes to detecting abnormalities in videos, the spatiotemporal nature of the video data makes it more challenging. This is because it requires processing a sequence of frames in a time-series format. Therefore, to overcome this challenge, it is crucial to extract significant features from every frame and consider their relationship with adjacent frames over time.

The associate editor coordinating the review of this manuscript and approving it for publication was Long Xu.

Convolutional Neural Networks (CNNs) have gained popularity in Deep Learning for extracting features from image data because they can learn hierarchical representations of image features. They extract in-depth features from high-dimensional data sets using complex structure and classification techniques, making them ideal for various applications [1]. Although CNNs are widely used in various deep learning tasks like text classification and Natural Language Processing (NLP) [2], they are mainly used in computer vision, like Face Recognition [3], Image Classification [4], and Object Detection [5].

On the other hand, Recurrent Neural Networks (RNNs) are known for their ability to model temporal dependencies in time-series data thanks to their ability to process information in both forward and backward directions. This allows the network to recall information from the past and use it to make informed decisions at the current time step. However, as information passes through multiple time steps, the data from the initial sequence may become diluted. To overcome this problem, advanced versions of RNNs, like long short-term memory (LSTM) [6] networks and gated recurrent units (GRUs) [7], have been developed, which can better retain information over more extended periods.

However, recent studies have suggested using deep learning architectures that combine CNNs and RNNs to enhance the performance of supervised models for violence detection in video data [8], [9]. This method efficiently extracts spatiotemporal characteristics by utilizing a CNN model to collect critical features from each video frame and then feeding them to an RNN model to analyze their temporal relationships and forecast whether any violent events happened in a video.

Aside from the mentioned methods, AI research has also concentrated on reducing the gap between human and machine behaviour in detecting violence through attention mechanisms. In computer vision, attention mechanisms were introduced to imitate the human visual system with a natural ability to find salient areas in complex scenes. Primarily, this can dynamically adjust the weight of input image features [10]. Attention mechanisms have demonstrated their effectiveness in many visual tasks such as image classification [11], object detection [12], and video understanding [13]. Different types of attention mechanisms have been proposed and utilized in VD, including Self-Attention [14], Multi-Head Self-Attention (MHSA) [15], and Convolutional Block Attention Module (CBAM) [16].

Over recent years, the safety of people has been a concern for various areas of the world. Due to the global economic crisis and the current socio-economic differences, the number of violent crimes has increased. Fig 1 presents the police-reported crime statistics for Canada between 2013 and 2021, clearly illustrating the upward trend in violent incidents during this period [30]. As a result, it underscores the critical need for systems that detect these violent crimes, thus contributing to a more secure society. The complexity of violence detection, which requires identifying anomalous events over time, is a primary focus of this study. The motivation for this paper is to develop a model that employs a novel methodology for detecting abnormal events, focusing on those anomalies that represent violent behaviour, given their significant implications for public safety and security.

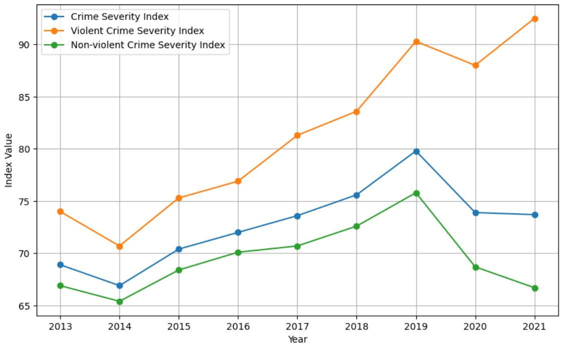

FIGURE 1. Crime severity index over the years [18].

This manuscript proposes a combination of a CNN (ResNet50) and an RNN (ConvLSTM) that utilizes MHSA modules [13] for VD from surveillance cameras. We will explore multiple models utilizing publicly available datasets, such as UCF-Crime [9] and RWF [17]. Subsequently, we will propose strategies to enhance the accuracy and robustness of these models in real-world scenarios.

The main contributions of this paper are as follows:

- We have designed a structure that uses an MHSA layer followed by a ConvLSTM cell, which brings the information of each attention layer to the next one.
- We have developed a unique VD model, KianNet, that merges MHSA-ConvLSTM with the ResNet50-ConvLSTM architecture for violence detection. This approach captures complex spatiotemporal features in videos, improving violence identification.
- We have comprehensively evaluated our model and showed that it outperforms other state-of-the-art algorithms.

The subsequent sections of this paper entail a comprehensive analysis of relevant works that employ distinct models, along with their respective sub-models, for detecting violence in surveillance cameras (Section II). After that, we present our proposed model (Section III) and evaluate its performance through several experiments (Section IV). Finally, we conclude with a discussion on future research ideas in Section V.

II. RELATED WORKS

The evolution of VD models has gained significant attention in recent years due to the increasing need for automated solutions to address violence in various settings. Researchers have proposed numerous approaches for VD, leveraging visual features extracted from video frames. This section will provide an overview of different approaches in VD.

A. 3D-CNN

Three-dimensional CNNs are a type of deep learning architecture used for video analysis tasks requiring spatial and temporal information. 3D-CNN is an extension of 2D-CNN that can handle video sequences as inputs. The 3D CNN architecture typically consists of multiple convolutional layers and pooling layers that learn to extract spatiotemporal features from video sequences. The output of the convolutional layers is then passed through fully connected layers and activation functions to make the final prediction. 3D CNNs have been successfully applied in various video analysis tasks, including action recognition, gesture recognition, and video-based violence detection. By leveraging spatial and temporal information, 3D CNNs can achieve state-of-the-art performance on these tasks, mainly when dealing with complex and dynamic videos. In a recent study, Tran et al. [19] proposed a 3D CNN model that achieved state-of-the-art performance on the Sports-1M [20] dataset, which contains many violent and non-violent videos.

Also, Sultani et al. in [9] introduced an approach based on Multiple Instance Learning (MIL) [21], using 3D Convolutional [22] features from various video segments to train a fully-connected neural network framework the model only with video-level labels. Then, a remarkable ranking loss algorithm was utilized to analyze the network's performance between the highest and lowest-scored instances for each positive (includes abnormal videos) and negative (includes normal videos) bag.

Recently, Magdy et al. in [23] proposed Violence 4D model for automatic VD from video datasets. Violence 4D is composed of three primary components, which Dense optical flow, ResNet50 and 4D residual blocks leverage the capabilities of four-dimensional convolution neural networks V4D CNN. Three other techniques [24], [25], and [26] are also introduced for the VD problem as the latest approaches so far, which all of them are based on 3D-CNN for the feature extraction part.

Another use of 3D-CNN can be seen in two-stream CNN as a deep learning architecture frequently used in VD tasks. This method became famous because of its ability to capture spatial and temporal information. This approach involves processing video frames using two separate streams - a spatial stream that extracts static appearance information from the frames and a temporal stream that captures the motion information. The spatial stream feeds each frame's raw RGB pixel values into a CNN architecture to extract appearance features. The temporal stream, on the other hand, computes optical flow from the frames and feeds them into a separate CNN to extract motion features. Finally, the output features from both streams are merged to make a final prediction. In a recent study, Pratama et al. [27] proposed a two-stream 3D CNN model that uses RGB and optical flow images for VD.

B. CNN-RNN

Many researchers believe more than extracting features with CNNs is needed for video data. They maintain that there is a need for adding RNNs to their model to consider the extracted features in a time interval. Therefore, they proposed CNN-RNN models for anomaly detection in video datasets. CNN-RNN models are a type of deep learning architecture used for video analysis tasks requiring spatial and temporal information. They are designed to combine the strengths of CNNs and RNNs to capture spatial and temporal features from video sequences. In hybrid CNN-RNN models, the CNN component extracts spatial features from individual frames, while the RNN component captures temporal dependencies between adjacent frames. The CNN component typically consists of several convolutional and pooling layers that learn to extract features from individual frames. The RNN component, on the other hand, takes the output of the CNN component and processes it through a series of recurrent layers that capture temporal dependencies between adjacent frames.

Vosta and Yow [8] proposed a hybrid CNN-RNN model that uses both CNNs and RNNs to extract spatial and temporal features from the video frames. Hybrid CNN-RNN models in video-based violence detection have improved performance compared to models that use only CNNs or RNNs. These models can effectively capture spatial and temporal features, leading to better detection of violent events in videos. Later, by replacing ConvLSTM with ConvGRU, another model called ConvGRU-CNN was introduced in [28] for VD.

Another CNN-RNN model in VD was proposed in [29], where authors added Bi-Directional LSTM to a CNN feature extraction model for real-time anomaly detection in surveillance cameras.

C. ATTENTION-BASED

Attention-based models are deep learning architectures that selectively focus on certain parts of the input data while ignoring others [30]. They are designed to improve the performance of neural networks by allowing them to weigh the importance of different input features selectively. In traditional neural networks, all input features are given equal importance, regardless of their relevance to the task. Attention-based models, however, assign different weights to input features based on their importance. This allows the model to selectively attend to the most informative parts of the input while ignoring irrelevant information. In video-based violence detection, attention-based models can help the network selectively focus on the most informative frames or regions within a frame, leading to better performance. For example, some approaches use spatial attention to focus on specific regions within a frame, while others use temporal attention to focus on specific frames within a video. Using attention mechanisms, video-based violence detection models can achieve higher accuracy while reducing the computational cost.

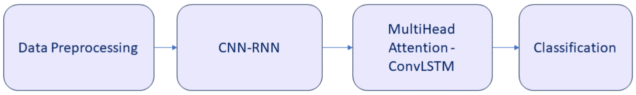

FIGURE 2. KianNet - the architecture.

In recent years several works have taken advantage of attention mechanisms for violence detection mainly in two categories of attention-based techniques: Self-Attention (citejoo2022clip,zhang2021generative) and MHSA (citezhou2023dual,rendon2021violencenet).

III. MODEL ARCHITECTURE

A. OVERALL ARCHITECTURE

The KianNet architecture has several steps, including Data preprocessing, CNN-RNN, MHSA-ConvLSTM, and Classification. Fig. 2 offers a general picture of the KianNet structure, showing how everything fits together.

- **Data preprocessing**: Each video file is divided into its frames in a desired format, and the difference between frame n and $n + 1$ will be calculated.
- **CNN-RNN**: The frame differences gained from the last step become an input for our ResNet50ConvLSTM structure to extract their features in a time series sequence.
- **MHSA-ConvLSTM**: The output of each ConvLSTM is fed to an MHSA layer to find the most important objects the model needs for detecting violence. This attention module is followed by another ConvLSTM cell to consider the recently extracted features in a sequence of frames.
- **Classification**: After passing through the final ConvLSTM layer, the output undergoes multiple max pooling and fully connected layers to determine if the input video is normal in binary classification. However, the model aims to identify the specific type of violent event in multi-class classification for each video input.

Fig. 3 shows the whole structure of KianNet in details of each primary steps; Data Preprocessing, CNN-RNN, MHSA-ConvLSTM, and Classification. Each of these main stages are discussed in the following sections:

B. DATA PREPROCESSING

Video cameras capture videos depending on their supporting resolution shown by FPS (Frame Per Second), which shows how many frames are captured in a second. For instance, a video that is recorded with an FPS of 30 for 10 seconds will comprise 300 frames (30 frames per second × 10 seconds = 300 frames). The figure in Fig. 4 depicts a selected set of frames from a video file that the model will be trained on.

In order to obtain a certain number of frames for the model, some frames must be skipped, with the number determined by *skipped_frames*. For instance, if we require 20 frames as input and the video file contains 300 frames, *skipped_frames* would equal 15. Thus, we will choose every 15th frame to create our input sequence frames. Additionally, it is important to consider that each video in the dataset has a different size. Therefore, every frame needs to be resized to a unique dimension to enhance compatibility with the model that processes the frames. For this research, we have resized each frame to a resolution of 224 × 224 pixels, which allows it to work seamlessly with the ResNet50 model.

Given that this work aims to detect violence in videos, we are not just looking at each frame as a standalone input. Instead, we use the difference between two consecutive frames to highlight the action. Therefore, we subtract each frame from the next one to gain the action between time t and $t + 1$ and ignore the parts that did not move during the time. Then, the produced image from subtraction frames will be the input for the feature extraction model. Instead of analyzing each video frame, we need to find the differences between frames in a period to show the movements. Figure 5 shows the difference between two consecutive frames. Since each frame is a uniform size of (224 × 224), and this size remains constant after subtraction, the input format for ResNet50 will be $(n\_frames, n\_row, n\_column, n\_channels)$, which in our case is $(20, 224, 224, 3)$.

C. CNN-RNN

1) CNN: RESNET50

Using CNN models in deep learning has become increasingly popular for extracting features from image data. These models are built with multiple layers, including convolutional and pooling layers, that can identify the most crucial features of input images. Several CNN architectures are developed each year to handle different subjects and datasets. Some of these models have become more widely used due to their exceptional performance and efficiency. Examples of such models include VGG, Inception, and ResNet, which have various versions available [34].

While a CNN structure with multiple layers can assist the model in identifying intricate features, the network's depth can cause vanishing gradient problems that can result in slow convergence or even halt the learning process [35]. One of the techniques to mitigate the vanishing gradient problem and improve the training of CNNs is using skip connection and residual blocks, which allows the gradients to flow more directly through the network, bypassing some layers and reducing the impact of the vanishing gradient problem [36].

In Fig. 6, we depict our experiments with different CNN structures in our model to find the best technique to provide us with higher accuracy with fewer parameters. The diagram shows that different models' parameters vary based on their network's depth and the convolutional layers they incorporate. However, accuracy only sometimes follows this trend. Although ResNet152 has more trainable parameters than ResNet50, it does not achieve higher accuracy on the UCF-Crime dataset. Despite having more parameters, this discrepancy can be attributed to factors like the potential

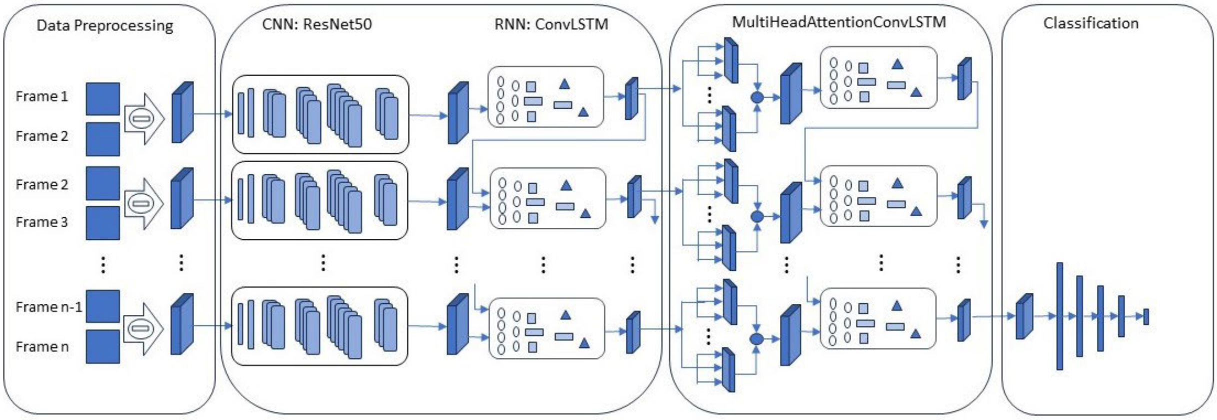

FIGURE 3. The detailed architecture of KianNet.

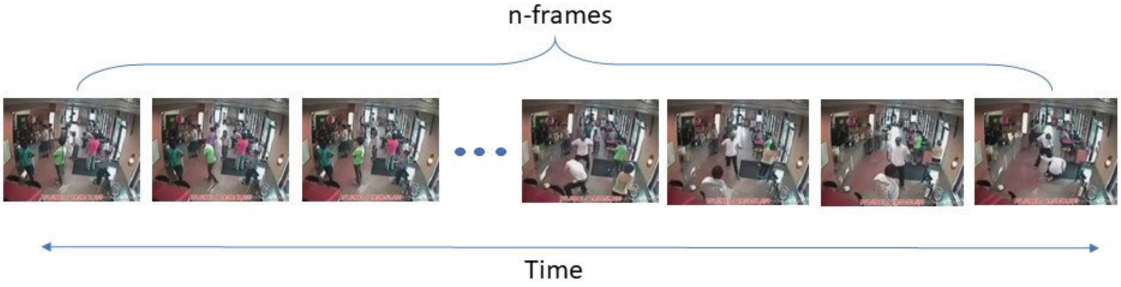

FIGURE 4. Divide a video file into its frames.

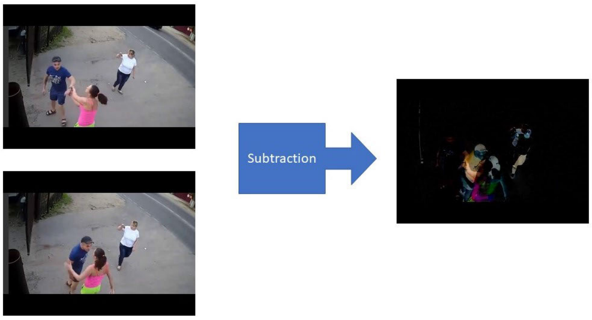

FIGURE 5. The subtraction of each neighbouring frame.

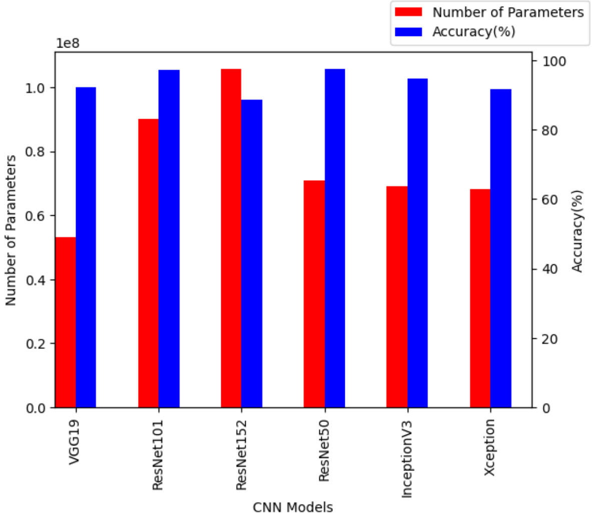

FIGURE 6. Comparison of using several CNN models in KianNet.

for overfitting, the specifics of the UCF-Crime dataset, and optimization challenges arising from issues such as vanishing or exploding gradients. As a result, we have chosen ResNet50 for KianNet as a CNN extraction because of its higher accuracy with fewer parameters. ResNet50 does have enough layers to extract features to detect the activities in a video and use residual blocks that prevent the model from the vanishing gradient problem [8].

Besides, ResNet50 has been pre-trained on large image datasets like ImageNet [37], which provides the model with a strong foundation for learning relevant features from images, including those related to violence. This pre-training enables the model to yield higher accuracy in violence detection tasks, especially when trained on limited datasets [38]. Fig. 7 also depicts the ResNet50 structure we used in our proposed model. This structure comprises five stages, each with a varying number of residual blocks, and each block consists of multiple convolutional layers. Besides, to better understand the shape of each layer's input and output, Table 1 provides the details of each step of ResNet50.

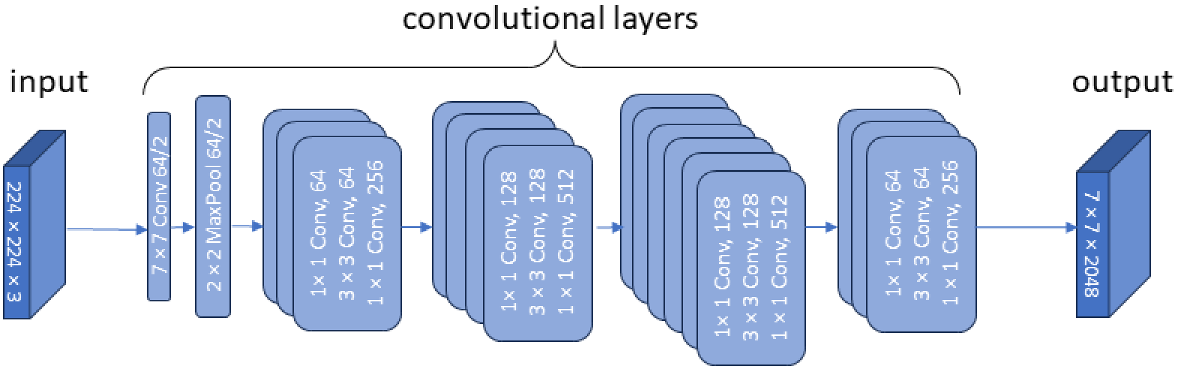

FIGURE 7. ResNet50 inner structure in our proposed model.

TABLE 1. The input and output size of each step in the proposed ResNet50.

Layer Type	Input Shape	Output Shape
Input Layer	(n,224,224,3)	(n,224,224,3)
Conv1 (7x7, stride 2)	(n,224,224,3)	(n,112,112,64)
MaxPooling (3x3, stride 2)	(n,112,112,64)	(n,56,56,64)
Conv Block 1 (3 layers)	(n,56,56,64)	(n,56,56,256)
Identity Block 1, 2	(n,56,56,256)	(n,56,56,256)
Conv Block 2 (3 layers)	(n,56,56,256)	(n,28,28,512)
Identity Block 3, 4, 5	(n,28,28,512)	(n,28,28,512)
Conv Block 3 (3 layers)	(n,28,28,512)	(n,14,14,1024)
Identity Block 6, 7, 8, 9, 10	(n,14,14,1024)	(n,14,14,1024)
Conv Block 4 (3 layers)	(n,14,14,1024)	(n,7,7,2048)
Identity Block 11, 12, 13	(n,7,7,2048)	(n,7,7,2048)

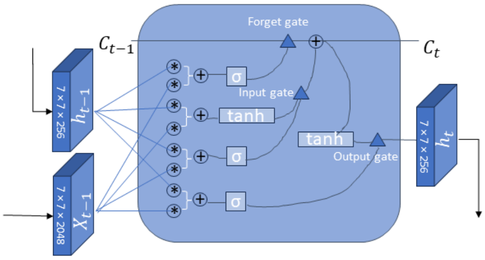

FIGURE 8. ConvLSTM operations in details.

2) RNN: CONVLSTM

Given that we work with video datasets composed of sequences of frames, we require a framework that can effectively handle time-series data. RNNs are renowned for managing time-series data in domains such as Natural Language Processing (NLP), speech recognition, and video analysis [39]. However, the standard RNNs also suffer from gradient vanishing problems like CNNs. To address this, units such as Long Short-Term Memory (LSTM) are invented to take advantage of having a "hidden state" in the network that can store information about the previous inputs, making them suitable for tasks requiring context or memory.

LSTM has become a valuable tool for handling time series data in neural network models. In this paper, we used the ConvLSTM model, one of the modified types of LSTM designed for dealing with images and works better than the standard vanilla LSTM. ConvLSTM is a neural network architecture that utilizes a convolutional layer at the input gate to extract spatial features from each sequence frame while capturing temporal dependencies between the frames using LSTM layers. This combination allows ConvLSTM to efficiently model the spatial-temporal structure in data, resulting in fewer parameters required for training [40]. Among Bi-LSTM, ConvLSTM, and ConvGRU, we chose ConvLSTM for our model. We preferred ConvLSTM over the other two because Bi-LSTM lacks a convolutional layer to handle spatio-temporal information, and ConvGRU does not have an explicit memory cell, which is crucial for capturing long-term information.

In KianNet's structure, the output of the ResNet50 from the previous stage has a size of $(n\_frames, 7, 7, 2048)$, which goes to the $X_{(t-1)}$ input of the ConvLSTM layer. We used a convolutional operation with 256 filters with a filter size of 3×3 and a stride of 1 in all the gates (input, forget, output, and the gate controlling the cell state). As a result, the hidden state of the ConvLSTM consists of 256 feature maps. This means that each of the gate mechanisms in the ConvLSTM operates in a convolutional manner, making this model particularly suited for tasks involving spatial data like images or video. The output (hidden state) is a three-dimensional tensor (for each step), maintaining the spatial structure of the input data while encoding temporal dependencies. Therefore, the output shape of our ConvLSTM will be $(n\_frames, 7, 7, 256)$.

D. MHSA-CONVLSTM

Researchers have been investigating methods to provide machines with consciousness to bridge the gap between humans and machines, thanks to the advancements in artificial intelligence over the past few decades [41], [42], [43]. One feature of human cognition that has been explored for implementation in machine learning approaches, particularly in computer vision, is the mental or vision saccades [44].

Many techniques in the field of computer vision have been proposed to integrate attention mechanisms into deep learning models. Self-Attention [30], MHSA [30], and the Convolutional Block Attention Module (CBAM) [16] are the most commonly used attention models. In our proposed VD, we used the MHSA technique to train the model to focus on specific points where violent events are more likely to occur. We analyzed the input feature map in a time sequence structure. We used the MHSA layer to enhance the accuracy of the final classification by selectively concentrating on crucial parts [45].

Our model uses the ConvLSTM output as an input (X) to the proposed attention module. As Equation 1 shows, each input X reshaped to the size of ($n\_row \times n\_column, n\_channels$), undergoes a transformation to provide Q, K, and V by using learned weight matrices W^Q, W^K, and W^V respectively.

$$
\begin{aligned}
Q &= XW^Q \\
K &= XW^K \\
V &= XW^V
\end{aligned} \tag{1}
$$

While the single attention function has d_{model}-dimensional keys, values, and queries, d_{model} for an MHSA layer will be h times d_k, d_k, and d_v of a set of queries, keys, and values. Then, these parameters are packed together into matrices Q, K, and V, respectively. The Attention function will be calculated as shown in Equation 2.

$$
Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}
$$

Since MHSA is composed of several single self-attention modules, and each head represents one scaled-dot attention layer, the MultiHead function concatenates the $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ as Equation 3 illustrates in the following

$$
MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O, \tag{3}
$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are weight matrices in this process.

Fig. 9 shows the details of the MHSA-ConvLSTM model used in KianNet, where we utilize the output from the ConvLSTM layer as input for our MHSA layer. This approach enables the model to concentrate on several objects, the same as the number of attention heads configured in the MHSA layer [46]. Following the application of these attention heads, the feature map for each input frame proceeds through another ConvLSTM layer.

The primary purpose of this step is to revisit and further process the features prioritized by the previous attention layers. Specifically, this second ConvLSTM layer facilitates the model's ability to consider these emphasized features again, but this time over a temporal sequence of frames. Therefore, the MHSA-ConvLSTM mechanism identifies the most important features within each frame and tracks and analyzes them across a series of frames. In designing KianNet, we strategically integrated the MHSA layer between two ConvLSTM layers as our model's specific components. The primary rationale behind this decision was to cater to scenarios where multiple objects are simultaneously involved in various types of violent behaviours. This integration allows our model to prioritize the most significant objects and analyze them in the context of their previous and subsequent frames. This distinctive configuration gives KianNet an edge over other architectures, improving its precision in detecting violent events.

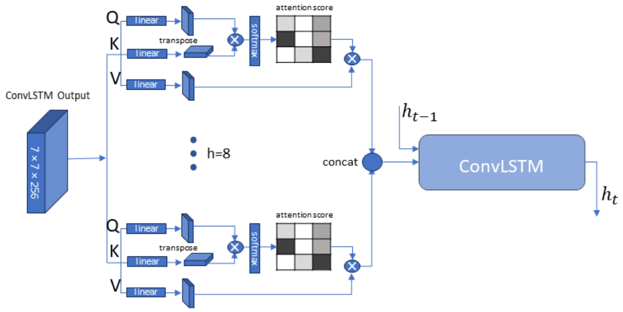

FIGURE 9. **The details of our proposed MHSA-ConvLSTM structure in KianNet.**

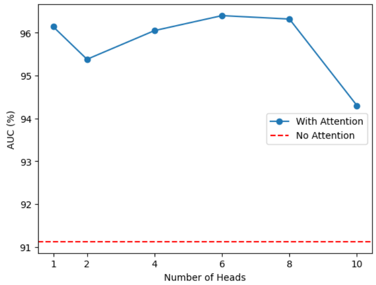

FIGURE 10. **Comparison the value of AUC in binary classification with different number of heads.**

One of the decisive factors in this attention technique is the number of heads, which shows the number of attention layers or heads used. Each head computes its attention scores, allowing the model to focus on different features in the input data. Each model can adjust the number of attention heads for their specific task depending on the dataset and techniques. Fig. 10 displays the value of AUC over the number of heads (h) in our experiments for our proposed model, KianNet, on the UCF-Crime dataset in binary classification. The blue line indicates the highest AUC value at 97.48% when $h = 8$. Consequently, we decided to use eight heads for our further experiments.

We use a multi-head self-attention layer because it can focus on several objects based on its number of heads. Although other methods like CBAM can be used in our model, the inner structure of our model, which contains two ConvLSTMs, provides us with convolutional layers with LSTMs, which work well on sequences of frames. This approach captures the spatial and temporal dynamics within

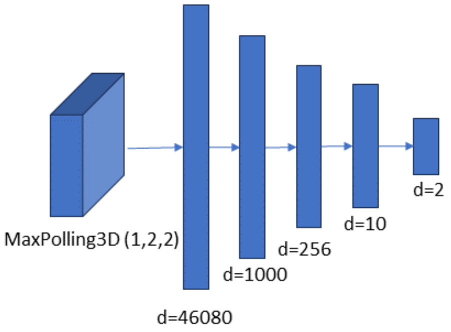

FIGURE 11. MaxPooling and Fully connected layers in classification stage of KianNet.

TABLE 2. Details for several datasets in violence detection.

Dataset	Data Scale	Length/Clip (sec)	Resolution
Hockey Fights [47]	1000 Clips	1.6-1.96	360 × 288
Movie Fights [47]	200 Clips	1.6-2	720 × 480
Crowd Violence [48]	246 Clips	1.04-6.52	Variable
UCF-Crime [9]	1900 Clips	60-600	Variable
RWF [17]	2000 Clips	5	Variable

the sequence, enhancing the model's overall understanding and interpretation of actions across time.

E. CLASSIFICATION

The final stage of the proposed model will be in the shape of a 4-dimensional tensor including *n_frames*, *n_rows*, and *n_column*, *n_channels*. Fig. 11 illustrates the layers of the classification stage. In that, after applying a 3D MaxPooling layer of size (2×2), we flatten the tensor (*n_frames*, 3, 3, 256) into a vector of the size of $(1, n\_frames \times 3 \times 3 \times 256)$ for classification object. Then, reduce the dimension by dropping out the less important features. Finally, we only need to use several fully connected layers of size 1000, 256, 10, *n_classes* to classify the input video as normal or violent (abnormal).

IV. EXPERIMENTS

In this section, we present the experimental results of our proposed model, "KianNet", and show how this model improves the violence detection performance in our trained video datasets, UCF-Crime [9] and RWF [17].

A. DATA

While finding a dataset that covers all types of violent behaviour may not be possible, several datasets shown in Table 2 can assist in violence detection.

One of the benchmark datasets in VD, which includes different types of violent behaviour captured by surveillance cameras, is UCF-Crime. Several examples of this dataset are shown in Fig. 12.

The UCF-Crime dataset includes 1900 videos of 13 crime categories captured from surveillance cameras in various situations with diverse backgrounds. As a result, a model trained on this dataset is more likely to detect violence when presented with new input videos accurately. UCF-Crime is usually considered a binary classification where it divides into a group of 950 data, including 13 types of crimes and 950 data for normal scenes. However, they can also be considered individual groups for detecting their specific type of crime. In this case, we only used 700 videos, including 50 from each of the 14 categories.

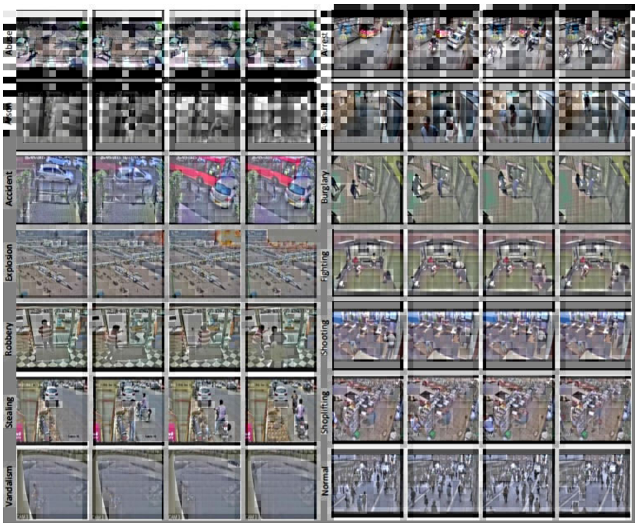

FIGURE 12. Example images of UCF-Crime dataset [9].

UCF-Crime has been modified in [8] by adding two new subcategories: 4MajCat and NREF.

4MajCat sub-dataset divides the dataset into four major categories: Theft, Vandalism, Violent behaviours, and Normal, which distinguishes more clearly between videos.

NREF contains 300 videos of Normal Road accidents, Explosions, and Fighting split into 5 seconds. In this sub-dataset, the Normal category is gained by the trimmed parts of the violent video, which has the same objects and backgrounds, which helps the model to be trained more accurately.

Table 3 explains the details of the UCF-Crime dataset when it is used for binary classification (Binary), all categories (AllCat), and the two modified sub-categories, 4MajCat and NREF.

Another dataset captured from real-world scenes using surveillance cameras is RWF, which contains 2000 real-world fighting videos for 5 seconds. Fig. 13 represents several samples of the RWF dataset.

We chose the two benchmark datasets, UCF-Crime and RWF because they derive from real-world events. In contrast to other datasets used in VD, such as HockeyFight, where data is collected in the same environments with many similar objects, UCF-Crime and RWF encompass a wide range of scenarios, positions, situations, and objects. As a result, our proposed model is more likely to perform efficiently in real-world applications when adequately trained on these datasets.

TABLE 3. Details of the UCF-Crime dataset's variants; Binary, AllCat, 4MajCat, and NREF.

Binary	No. Videos	AllCat	No. Videos	4MajCat	No. Videos	NREF	No. Videos
Abuse	50	Abuse	50	Theft (Burglary, Robbery, Shoplifting, Stealing)	150	RoadAccident	30
Arrest	50	Arrest	50				
Arson	50	Arson	50				
Assault	50	Assault	50				
Burglary	100	Burglary	50	Vandalism (Arson, Explosion, RoadAccident, Vandalism)	150	Explosion	50
Explosion	50	Explosion	50				
Fighting	50	Fighting	50				
RoadAccident	150	RoadAccident	50				
Robbery	150	Robbery	50	Violence behaviours (Abuse, Arrest,Assault, Fighting, Shooting)	150	Fighting	70
Shooting	50	Shooting	50				
Shoplifting	50	Shoplifting	50				
Stealing	100	Stealing	50				
Vandalism	50	Vandalism	50	Normal	150	Normal	150
Normal	950	Normal	50				

FIGURE 13. Example images of RWF dataset [17].

In the next section, we will evaluate our proposed method using the aforementioned datasets and compare its performance with other models to demonstrate its efficacy in different approaches.

B. PERFORMANCE METRICS

This study uses two primary metrics to evaluate the model's performance: Accuracy and AUC (Area Under Curve). Accuracy measures the percentage of correctly classified instances by a model (Equation 4). It is calculated as the ratio of correctly classified instances to the total number of samples in the dataset. In the case of binary classification, accuracy can be calculated as Equation 5, where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

$$Accuracy = \frac{\#Correct\_predictions}{\#Total\_predictions} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

While accuracy can be a helpful metric for evaluating the overall performance of a model, it can be misleading in some cases. For instance, in a dataset with imbalanced classes, where one class is much more prevalent than the other, a model that predicts the majority class for every instance can achieve high accuracy, even if it fails to classify instances of the minority class correctly. Another drawback of using accuracy is using a threshold (default set to 0.5 for binary classification), which is tricky. If the threshold is set too low, the model classifies many scenes as violent files, including many normal activities. Also, if set too high, many crimes will be missed, increasing the false negative rate.

On the other hand, AUC is a performance metric used to evaluate the quality of a binary classification model's predictions. It measures the ability of the model to distinguish between the normal and abnormal classes in our case. AUC is typically used in the context of the receiver operating characteristic (ROC) curve, which is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. Therefore, AUC can work as a threshold-independent metric. The AUC is then calculated by computing the area under the ROC curve, which ranges from 0 to 1, in that, as the value is close to 1, it shows a more robust classifier.

C. SETTINGS

We did several experiments for our proposed model, KianNet, using ResNet50, ConvLSTM, and MHSA via *Keras* libraries. We set $batch\_size = 16$, $learning\_rate = 1 \times 10^{-4}$, $epochs = 50$, *glorot_uniform* as the initial weight, and *RMSprop* as an optimizer to compare the KianNet model with other methods on UCF-Crime and RWF datasets.

D. EVALUATION AND COMPARISON

1) EXPERIMENTS ON RWF

RWF has become one of the benchmark datasets in recent years. Several models achieved high accuracy, more than 80% in classifying violent or non-violent behaviours like 2D

TABLE 4. Binary classification on RWF dataset based on accuracy.

Author(s)	Model	Accuracy (%)
Sudhakaran et al. [40]	Convolutional LSTM	77
Tran et al. [22]	C3D	82.75
Cheng et al. [16]	Flow Gated Net	87.25
Su et al. [50]	SPIL Convolution	89.3
Islam et al. [51]	SepConvLSTM-M	89.75
Pratama et al. [27]	Two-stream 3D CNN	90.50
Kang et al. [52]	2D CNNs + LSTM	92
Chelali et al. [53]	2D Spatio-Temporal	93.80
Magdy et al. [23]	Violence 4D	94.67
Proposed method	KianNet	**96.21**

TABLE 5. Binary classification on UCF-Crime dataset based on AUC.

Author(s)	Model	AUC (%)
Sultani et al. [9]	SVM	50
Tur et al. [24]	k-diffusion	65.22
Simonyan et al. [54]	VGG-16	72.66
Liu et al. [25]	PFMF	74
biradar et al. [55]	DEARESt	76.66
Zhong et al. [56]	TSN-OpticalFlow	78.08
Zhong et al. [56]	C3D	81.08
Vosta et al. [8]	ResNetConvLSTM	81.71
Qasim et al. [28]	ConvGRU-CNN	82.65
Tian et al. [57]	RTFM	84.30
Ullah et al. [29]	Multi-layer BD-LSTM	85.53
Sun et al. [26]	LSTC	85.88
Zhou et al. [33]	UR-DMU	86.97
Joo et al. [31]	CLIP-TSA	87.58
Proposed method	KianNet	**97.48**

CNN+LSTM [49] and Violence 4D [47]. We run KianNet on the RWF dataset to evaluate the capability of our model in violence detection with real-world fighting movies. In our experiments, we trained the model by 80% of the dataset while the rest was selected for testing. Table 4 compares KianNet with several other models on the RWF dataset based on accuracy, where our proposed model achieved the highest accuracy of 96.21% for violence detection.

2) EXPERIMENTS ON UCF-CRIME

The UCF-Crime dataset includes 13 types of anomalies, while the rest are all normal scenes. Many researchers evaluated their model using AUC for their experiments on the UCF-Crime dataset. This is because AUC is a suitable performance metric due to its threshold independence feature and the ability to work with imbalanced data, which can play an essential role in UCF-Crime multi-class classification tasks, where each category has various samples.

In Table 5, several VD models are compared using AUC in binary classification on the UCF-Crime dataset. As we can see from Table 5, the MIL-C3D model proposed by Sultani et al. in their paper [9] gained 74% in AUC. Also, Zhong et al. in [17] presented TSN models based on RGB and optical flow, with the value of AUC 82% and 78%, respectively. However, one of the best models for violence detection on UCF-Crime was proposed by Ullah et al. in [29] where they applied a multi-layer BD-LSTM technique to achieve 85% in AUC.

TABLE 6. Precision, recall, and F1-score equations.

Precision	$\frac{\#TruePositives}{\#TruePositives+\#FalsePositives}$
Recall	$\frac{\#TruePositives}{\#TruePositives+\#FalseNegatives}$
F1-Score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

TABLE 7. Comparison between KianNet and ResNetConvLSTM on the UCF-Crime AllCat dataset.

Model	Precision	Recall	F1-Score
ResNetConvLSTM	22.93	24.31	23.60
KianNet	**24.23**	**25**	**24.60**

TABLE 8. Ablation study of architecture ResNetConvLSTM and KianNet: ResNet50ConvLSTM-MHSA-ConvLSTM on UCF-Crime original and modeified datasets based on accuracy and AUC.

Model	ResNetConvLSTM		KianNet	
Dataset	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)
NREF	79.04	65.38	83.14	73.84
4MajCat	73.88	62.22	88.91	73.75
AllCat	53.88	22.72	63.71	23.88
Binary	81.71	62.50	97.48	92.98

E. ABLATION STUDY

For the ablation study, a double experiment was proposed to test how using a multi-head self-attention module followed by a ConvLSTM layer mechanism affected the violence detection on the UCF-Crime dataset regarding Accuracy and AUC. Although a more powerful backbone network was used than in previous work, we considered it interesting to check how the performance improved by using the attention mechanism. When comparing our proposed model with the one without the MHSA-ConvLSTM module, we obtained better results in accuracy and AUC. Table 8 presents the results of the ablation study of the architecture of ResNet50ConvLSTM and KianNet, where the MHSA-ConvLSTM module was added to the previous model. As can be seen in Table 8, both accuracy and AUC were consistently better when the attention module was used. The most significant improvement of using KianNet happened when the model was applied to the binary classification dataset, where the AUC value rose from 81.71 to 97.48 percent. There are also improvements in violence detection performance in other datasets, NREF, 4MajCat, and AllCat. Another challenge in UCF-Crime is classifying each video in the exact match class in AllCat. Therefore, the classifier should classify each input into one of the 13 crime types and normal in AllCat. The situation worsens when some categories are too similar to distinguish them, like shoplifting and stealing. However, KianNet improved the accuracy value for this classification marginally from 22.72% to 23.88%.

In Table 7, we also compared KianNet with ResNetConvLSTM [8] in Precision, Recall, and F1-Score in addition to Accuracy and AUC to show the importance of the added attention mechanism to our proposed model. Table 6 shows how each measurement will be calculated. Precision calculates by aiming to diminish the effect of False Positive.

Recall is vital because of the ability to show the importance of False Negatives, which is essential in violence detection problems. Considering both values, the F-1 score is the harmonic mean of Precision and Recall. Therefore, we have a much more reliable model when the value of the F1-Score is higher.

V. CONCLUSION AND FUTURE WORK

This paper introduced KianNet, an approach for violence detection from surveillance camera footage. To deal with such video datasets, we used ResNet50 to extract features from each video frame and the ConvLSTM technique for considering the relationship between frame sequences. We also brought vision saccade to our model through MHSA to make the model more conscious, like how the human brain works. We conducted extensive experiments using the UCF-Crime dataset (original and modified versions) and the RWF dataset to test our proposed model, KianNet. The results demonstrated KianNet's superior performance to other violence detection techniques in binary classification. This further underlines the potential of our approach for practical implementations in violence detection and prevention.

- Despite the outstanding performance of KianNet in Violence Detection, the number of training parameters is high due to the use of two attention mechanisms in the model's architecture. Therefore, we will work on designing a lightweight VD attention mechanism in our future work.
- To better understand the actions happening in a video file, we can offer a technique to recognize the action after the feature extraction part by using YOLOv3 to recognize the extracted body part and then build separate ConvLSTM to learn the movement patterns of each body part.
- KianNet can also be applied to other areas to analyze and detect several events. With its unique learning structure and strong performance in detecting violent behaviour from video surveillance, it can be effectively employed in areas such as fall detection in homecare settings or hospitals.
- Another improvement we can make to our technique is using the original image alongside the moving parts gained from the subtraction of frames to improve the feature extraction.
- Since we work on videos, which usually have sounds, it would be much more helpful if we use the sounds as a separate line of input to the model to detect violent actions in videos more accurately.

REFERENCES

[1] J. Lu, L. Tan, and H. Jiang, "Review on convolutional neural network (CNN) applied to plant leaf disease classification," *Agriculture*, vol. 11, no. 8, p. 707, Jul. 2021.

[2] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1150.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[7] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2017, pp. 1597–1600.

[8] S. Vosta and K.-C. Yow, "A CNN-RNN combined structure for real-world violence detection in surveillance cameras," *Appl. Sci.*, vol. 12, no. 3, p. 1021, Jan. 2022.

[9] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[10] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[11] Y. Chen, D. Zhao, L. Lv, and C. Li, "A visual attention based convolutional neural network for image classification," in *Proc. 12th World Congr. Intell. Control Autom. (WCICA)*, Jun. 2016, pp. 764–769.

[12] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[13] F. J. Rendón-Segador, J. A. Álvarez-García, F. Enríquez, and O. Deniz, "ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence," *Electronics*, vol. 10, no. 13, p. 1601, Jul. 2021.

[14] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, Apr. 2020.

[15] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107676.

[16] B. Chen, Z. Zhang, N. Liu, Y. Tan, X. Liu, and T. Chen, "Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition," *Information*, vol. 11, no. 8, p. 380, Jul. 2020.

[17] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4183–4190.

[18] G. Moreau, "Police-reported crime statistics in Canada," Uniform Crime Reporting Surv. (UCR), Canada, Tech. Rep. 3302, 2022.

[19] J. Zhang and Z. Liu, "Detecting abnormal motion of pedestrian in video," in *Proc. Int. Conf. Inf. Autom.*, Jun. 2008, pp. 81–85.

[20] J. Zhang and Z. J. Liu, "Abnormal behavior of pedestrian detection based on fuzzy theory," Deakin Univ., Australia, Tech. Rep., 2023.

[21] G. Liu, J. Wu, and Z.-H. Zhou, "Key instance detection in multi-instance learning," in *Proc. Asian Conf. Mach. Learn.*, 2012, pp. 253–268.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[23] M. Magdy, M. W. Fakhr, and F. A. Maghraby, "Violence 4D: Violence detection in surveillance using 4D convolutional neural networks," *IET Comput. Vis.*, vol. 17, no. 3, pp. 282–294, Apr. 2023.

[24] A. O. Tur, N. Dall'Asen, C. Beyan, and E. Ricci, "Exploring diffusion models for unsupervised video anomaly detection," 2023, *arXiv:2304.05841*.

[25] Z. Liu, X.-M. Wu, D. Zheng, K.-Y. Lin, and W.-S. Zheng, "Generating anomalies for video anomaly detection with prompt-based feature mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24500–24510.

[26] S. Sun and X. Gong, "Long-short temporal co-teaching for weakly supervised video anomaly detection," 2023, *arXiv:2303.18044*.

[27] R. A. Pratama, N. Yudistira, and F. A. Bachtiar, "Violence recognition on videos using two-stream 3D CNN with custom spatiotemporal crop," *Multimedia Tools Appl.*, vol. 82, pp. 1–23, 2023.

[28] M. Q. Gandapur and E. Verdú, "ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 8, no. 4, p. 88, 2023.

[29] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 16979–16995, May 2021.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[31] H. Kevin Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection," 2022, *arXiv:2212.05136*.

[32] W. Zhang, G. Wang, M. Huang, H. Wang, and S. Wen, "Generative adversarial networks for abnormal event detection in videos based on self-attention mechanism," *IEEE Access*, vol. 9, pp. 124847–124860, 2021.

[33] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," 2023, *arXiv:2302.05160*.

[34] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," *Frontiers Neurosci.*, vol. 13, p. 95, Mar. 2019.

[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, in Proceedings of Machine Learning Research, vol. 9, Y. W. Teh and M. Titterington, Eds., Sardinia, Italy, May 2010, pp. 249–256.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4278–4284.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[38] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, no. 7, pp. 1851–1864, Jul. 2013.

[39] G. Zhou and Y. Wu, "Anomalous event detection based on self-organizing map for supermarket monitoring," in *Proc. Int. Conf. Inf. Eng. Comput. Sci.*, Dec. 2009, pp. 1–4.

[40] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[41] Y.-p. Tang, X.-j. Wang, and H.-f. Lu, "Intelligent video analysis technology for elevator cage abnormality detection in computer vision," in *Proc. 4th Int. Conf. Comput. Sci. Converg. Inf. Technol.*, Nov. 2009, pp. 1252–1258.

[42] J. Feng, C. Zhang, and P. Hao, "Online learning with self-organizing maps for anomaly detection in crowd scenes," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3599–3602.

[43] M. H. Sharif, S. Uyaver, and C. Djeraba, "Crowd behavior surveillance using Bhattacharyya distance metric," in *Proc. Int. Symp. Comput. Model. Objects Represented Images*. Cham, Switzerland: Springer, 2010, pp. 311–323.

[44] Oleg Gorokhov, Mikhail Petrovskiy, and Igor Mashechkin, "Convolutional neural networks for unsupervised anomaly detection in text data," in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 500–507. Springer, 2017.

[45] B. Ramachandra and M. J. Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2558–2567.

[46] X. Wen and W. Li, "Time series prediction based on LSTM-attention-LSTM model," *IEEE Access*, vol. 11, pp. 48322–48331, 2023.

[47] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. 14th Int. Conf. Comput. Anal. Images Patterns (CAIP)*, Seville, Spain. Berlin, Germany: Springer, Aug. 2011, pp. 332–339.

[48] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–6.

[49] B. Zapata-Impata, P. Gil, and F. Torres, "Learning spatio temporal tactile features with a ConvLSTM for the direction of slip detection," *Sensors*, vol. 19, no. 3, p. 523, Jan. 2019.

[50] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human interaction learning on 3D skeleton point clouds for video violence recognition," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 74–90.

[51] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient two-stream network for violence detection using separable convolutional LSTM," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[52] M.-S. Kang, R.-H. Park, and H.-M. Park, "Efficient spatio-temporal modeling methods for real-time violence recognition," *IEEE Access*, vol. 9, pp. 76270–76285, 2021.

[53] M. Chelali, C. Kurtz, and N. Vincent, "Violence detection from video under 2D spatio-temporal representations," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2593–2597.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] K. Biradar, S. Dube, and S. K. Vipparthi, "DEARESt: Deep convolutional aberrant behavior detection in real-world scenarios," in *Proc. IEEE 13th Int. Conf. Ind. Inf. Syst. (ICIIS)*, Dec. 2018, pp. 163–167.

[56] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[57] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4955–4966.

SOHEIL VOSTA (Student Member, IEEE) received the B.Sc. degree in computer science from the University of Isfahan, Iran, in 2015, and the M.Sc. degree in computer science-computational theory from Tarbiat Modares University, Iran. He is currently pursuing the Ph.D. degree in software system engineering with the University of Regina, Canada. His research interests include dimension reduction methods for image processing models and continued in deep learning and artificial intelligence techniques in video analysis. He is an active graduate student member for three years and an ExCom Member of the Region-7 South Saskatchewan Section.

KIN-CHOONG YOW (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the National University of Singapore, in 1993, and the Ph.D. degree from the University of Cambridge, U.K., in 1998. In September 2018, he joined the University of Regina, where he is currently a Professor with the Faculty of Engineering and Applied Science. Prior to joining the University of Regina, he was an Associate Professor with the Gwangju Institute of Science and Technology (GIST), Republic of Korea, from 2013 to 2018; a Professor with the Shenzhen Institutes of Advanced Technology (SIAT), China, from 2012 to 2013; and an Associate Professor with Nanyang Technological University (NTU), Singapore, from 1998 to 2013, where he was the Sub-Dean of Computer Engineering, from 1999 to 2005. He was the Associate Dean of Admissions with NTU, from 2006 to 2008. He has published more than 100 top quality international journal articles and conference papers. His research interests include artificial general intelligence and smart environments. He is a member of APEGS and ACM. He has served as a Reviewer for a number of premier journals and conferences, including IEEE Wireless Communications and IEEE Transactions on Education. He has been invited to give presentations at various scientific meetings and workshops, such as ACIRS, from 2018 to 2019; ICSPIC, in 2018; and ICATME, in 2021. He is the Editor-in-Chief of the *Journal of Advances in Information Technology* (JAIT).