IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 March 2024, accepted 18 March 2024, date of publication 21 March 2024, date of current version 28 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380192

RESEARCH ARTICLE

VD-Net: An Edge Vision-Based Surveillance System for Violence Detection

MUSTAQEEM KHAN[1], (Member, IEEE), ABDULMOTALEB EL SADDIK[1,2], (Fellow, IEEE), WAIL GUEAIEB[2], (Senior Member, IEEE), GIULIA DE MASI[3], (Senior Member, IEEE), AND FAKHRI KARRAY[1,4], (Fellow, IEEE)

[1]Department of Computer Vision, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[3]Technology Innovation Institute (TII), Abu Dhabi, United Arab Emirates
[4]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Corresponding author: Mustaqeem Khan (Mustaqeem.khan@mbzuai.ac.ae)

This work was supported in part by the Project "Intelligent Object Detection, Dynamic Scene and Activity Recognition for Real-Time Unmanned Aerial Vehicle (UAV) Applications," developed at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), funded by the Technology Innovation Institute (TII), Abu Dhabi, United Arab Emirates.

ABSTRACT The automation of surveillance systems, driven by the rapid development of computer vision technology, has significantly enhanced the analysis of surveillance videos, particularly in recognition of human activity, including behavior analysis and violence detection, thereby bolstering public and industrial security. Despite these advancements, detecting and analyzing violent actions remains challenging, especially for real-time surveillance systems with limited computing power. We propose an artificial intelligence-based framework called VD-Net (Violence Detection Network), enabled by Intelligent Internet-of-Things (IIoT) to detect violent behavior in public and private spaces. The model utilizes lightweight special task temporal convolutional network (ST-TCN) blocks and several bottleneck layers to focus on salient features in the input sequence. The learned features passed from the classifier to discriminate between violent and nonviolent actions. Additionally, our system is supposed to trigger an alert if violence is detected, which is then communicated to relevant departments. We checked the robustness of our system by surveillance and non-surveillance datasets and ensured a 1-4 % improvement in State-of-The-Art (SoTA) accuracy.

INDEX TERMS Artificial intelligence, cloud computing, edge intelligence, Internet of Things (IoT), security, smart city, violence detection.

I. INTRODUCTION

Technologies exist to automatically detect and flag violent behavior in various digital media formats, such as images, videos, audio recordings, and text [1]. The growing volume of digital content and the need to moderate user-generated content has increased interest in the technology [2]. Violence detection technology aims to promptly identify and eliminate violent content from online platforms, protecting users from exposure to potentially harmful. Intelligent surveillance systems analyze video patterns to detect violence in heavily populated areas for public safety in smart cities [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

The Intelligent Internet of Things (IIoT) can effectively detect and prevent online violence to promote safety. According to a report [3], 48% of deaths in 2015 were attributed to interpersonal violence, and 25% of happened with a sharp instrument, such as knives and razors. Similarly, there were approximately more than a million violent incidents in 2019 at the United States, including fights, aggressive behavior, and mass shootings. Hence, surveillance cameras are now widely used for automatic monitoring systems in both public and private sectors, but detecting abnormal activities with high accuracy remains a challenge [4]. However, replacing labor-intensive and tedious manual surveillance systems with automatic systems would further increase public and private security and safety levels. These technologies can

help law enforcement agencies detect suspected actions and differentiate between normal and abnormal behaviors.

Researchers from different fields are looking and carefully studying surveillance videos for new methods to recognize different actions and activities [5]. However, surveillance videos can be analyzed using activity recognition, summarization, and individual identification to identify specific activities [6]. Hence, violence detection is still missing in commercial and industrial monitoring systems due to the use of handcrafted and traditional neural network features. Implementing violence detection techniques in surveillance for real-time application using IIoT has remained a challenge.

In this manner, IIoT-based networks linked to vision sensors could aid law enforcement agencies in preventing crime in smart cities by analyzing input for crucial info like individuals/suspicious things. In such a system, IoT devices share data upon discovering violent things to examine the things and address safety concerns. Implementing this arrangement could allow for automatic monitoring of activity captured by multiple location cameras for managing unusual events. Hence, various algorithms have been developed for violence detection in monitored and unmonitored environments using traditional and deep learning-based methods [7].

However, these SoTA approaches face challenges such as high computational requirements, restricted viewpoints, varying lighting conditions, complex crowd scenarios, and changes in intensity. Furthermore, some limitations in this domain are false positives and negatives, limited detection scope, context dependency, and potential biases and unfairness, especially when models are trained on imbalanced datasets or datasets that contain stereotypes and prejudiced information. For these reasons, we develop a system that avoids perpetuating harmful stereotypes and biases and improves safety and security in public and private spaces. By providing real-time detection, enhanced situational awareness, improved accuracy, scalability, and integration with other systems, an IoT-based system can help prevent violent incidents and reduce the risk of harm to individuals and communities. The proposed VD-Net detects violence in public, private, and industrial settings by combining practical edge computing with cloud servers through three connected surveillance cameras. Considering the challenges and limitations mentioned above, this system will be precious. The main contributions are as follows:

- Traditional monitoring systems often have wire failures during installation, resulting in slow response times and increased processing requirements for authorities. To tackle this challenge, we propose an AI-driven framework for violence detection that leverages the powerful capabilities of Internet-of-Things (IoT) to connect devices for the smooth exchange of information. Moreover, we develop a cloud-based system that enables comprehensive investigations of violent incidents in public and private settings with fast processing.
- To process surveillance data, we need an intelligent edge-based mechanism to extract helpful information during analysis. To tackle this issue, we investigate the use of IoT and introduce a lightweight system that can be implemented on embedded devices. Our system recognizes critical violence to process and transmit over the network for detailed investigation in the cloud instead of all frames. This approach streamlines the process and improves its overall intelligence.
- Traditional IoT methods often use manual features or clustering algorithms, which may not capture long-term dependencies, reducing accuracy. We propose a bottleneck layer in VD-Net to encode spatial and temporal correlations and analyze local motion between frames to overcome this issue. Additionally, the cloud server acquires feature vectors and sends them to an attention unit for identifying salient cues. This is the first time bottleneck layers are utilized for violence, significantly improving accuracy with reduced latency for real-time applications.
- To the best of our knowledge, this article represents the first use of a bottlenecks layer to learn salient cues of violent activity in the IIoT network. The module extracts information from the input layer and determines whether a scene is violent or nonviolent in a public/private environment. We evaluated the proposed VD-Net using publicly available datasets demonstrating outperformed SOTA approaches. Additionally, the system can be used for indoor and outdoor surveillance in IIoT-based systems.

The article are organized as follows. Section II presents a comprehensive overview of the relevant literature pertaining to the topic under investigation. Section III elucidates the proposed system in meticulous detail. Section IV delineates the empirical findings and their corresponding analysis, supplemented by a comparative study. Ultimately, Section V encapsulates the conclusions derived from this research endeavor and posits potential avenues for future directions.

II. RELATED WORK

Overall, violence detection is a complex and challenging research area. Researchers are constantly exploring new techniques to advance the surveillance system for public safety. Ethical considerations surrounding privacy and potential biases in violence detection (VD) algorithms must also be considered in developing and deploying violence detection systems. In contrast, we cover the latest advances in this field, including conventional and deep learning-based techniques that have drawn much interest within the research community.

A. MACHINE LEARNING-BASED VD APPROACHES

This section provides an overview of key research conducted in machine learning for violence detection. One of the early works in this area [8] developed a machine learning-based

approach to detect violent movie scenes. The authors used a set of visual and audio features to classify scenes as violent or nonviolent and achieved an 85 % accuracy on the movie fight dataset. Similarly, in [9], the authors presented a machine learning-based approach to detect violent events in surveillance videos using handcrafted features, such as motion and texture, to classify the violence. Furthermore, the authors in [10] introduced violence detection for social media and used a set of visual and audio features to classify the actions accurately. However, [11] and [12] endorsed a new machine learning-based approach to detect violent movie events.

Similarly, a conventional method was proposed in [13], utilizing motion cues derived from optical flow using RGB frames and incorporating appearance as low-level features. By eliminating redundant information, the system developed a bag of words (BoWs). Similarly, [14] developed a system to identify violence in crowded settings based on background motion correction, appearance, and long-term dependencies. In order to demonstrate how violent events are related to scene-scale spatial events, they used late fusion and BoW. Another approach [15] developed a new local descriptor to manage and reduce the coefficient reconstruction error to present a sparse-based model for classification. Furthermore, [16] incorporated pixel-based analysis and object trajectory results to monitor object speed, direction, and smaller movements. Hence, the practice of these methods grew tiresome due to hand-carried engineering. The subsequent section provides an overview of advance techniques.

B. DEEP LEARNING-BASED VD APPROACHES

Recently, deep learning techniques have become more popular for detecting violence. The early works in this area [17] presented a method based on deep learning to identify instances of violent behavior in surveillance videos. The authors used a two-stream convolution neural network (CNN) architecture to extract spatiotemporal features from the videos and achieved 89 % classification accuracy on the surveillance dataset. Similarly, in [18], the authors developed a deep learning-based model for violence detection in social media and extracted visual features and temporal dependencies by a long-short term memory (LSTM) and achieved 94.9 % results on the same dataset. Moreover, the authors in [19] and [20] introduced a deep learning approach for detecting violent events in urban surveillance videos. Furthermore, [21] and [22] presented a new method to detect abnormal events in movie datasets.

Computer vision challenges are being addressed with deep learning based on recent studies. However, there are also concerns that such technology is being used for violence. For instance, a method in [7] represents a frame in a sequence using critical information provided by Hough's feature. Liu et al. [23] utilized a 3D CNN to identify violent scenes in video-applied sampling as a pre-processing step. The researchers developed a deep learning-based model for detecting violent scenes utilizing transfer learning techniques, while [24] introduced a Spark framework for detecting violent scenes by bidirectional LSTM. Similarly, [25] introduced the idea of aggregating the ensembles, and [26] employed a combination of 3-D CNN and support vector machine (SVM) to identify violent actions in videos. However, a comprehensive literature analysis indicates that many existing methods must be revised to overcome several limitations and challenges. These include inadequate integration with state-of-the-art IoT devices, heavy reliance on end-to-end pre-trained models, failure to incorporate cloud-based concepts, and the use of handcrafted features.

C. ATTENTION-BASED VD APPROACHES

There has been a growing interest in using attention and transformer techniques for violence detection in various contexts, such as social media and surveillance videos. A recent work [27] presented a model based on the attention mechanism to detect violent movie scenes. The authors used a two-layered LSTM network with an attention mechanism to identify violent scenes in the movie fight dataset. Similarly, in [28], the authors introduced a transformer-based model for violence detection in social media and used a multiheaded self-attention mechanism to capture the temporal relationships between frames and a global temporal encoding layer to aggregate the features representations. Another study [29] developed a transformer-based model for detecting violent events in online streaming and used a hierarchical attention mechanism to capture the spatial and temporal relationships in-game violent scenes on a gaming dataset.

Similarly, [30] introduced a model that uses a multiheaded self-attention mechanism to detect violent content in surveillance videos. Hence, [31] introduced a model based on the transformer architecture with a dual-branch structure to detect violent events. They processed the frame-level and shot-level features and achieved high accuracy on surveillance datasets. In conclusion, attention and transformer techniques have shown promising results for violence detection in various contexts, including social media, surveillance videos, and online gaming platforms. In this regard, the proposed model uses attention mechanisms to capture the temporal and spatial relationships between sequences and has achieved high accuracy on several datasets.

III. METHODOLOGY

This section thoroughly explains the proposed violence detection network by examining each module, as illustrated in Figure 1. The authors in [32] developed a violence detection system for industrial settings and covered indoor valences through an advanced IIOT system. This method has some limitations, such as covered indoor activities, computational complexity, and latency for real-time applications. We were motivated by their work and proposed an advanced

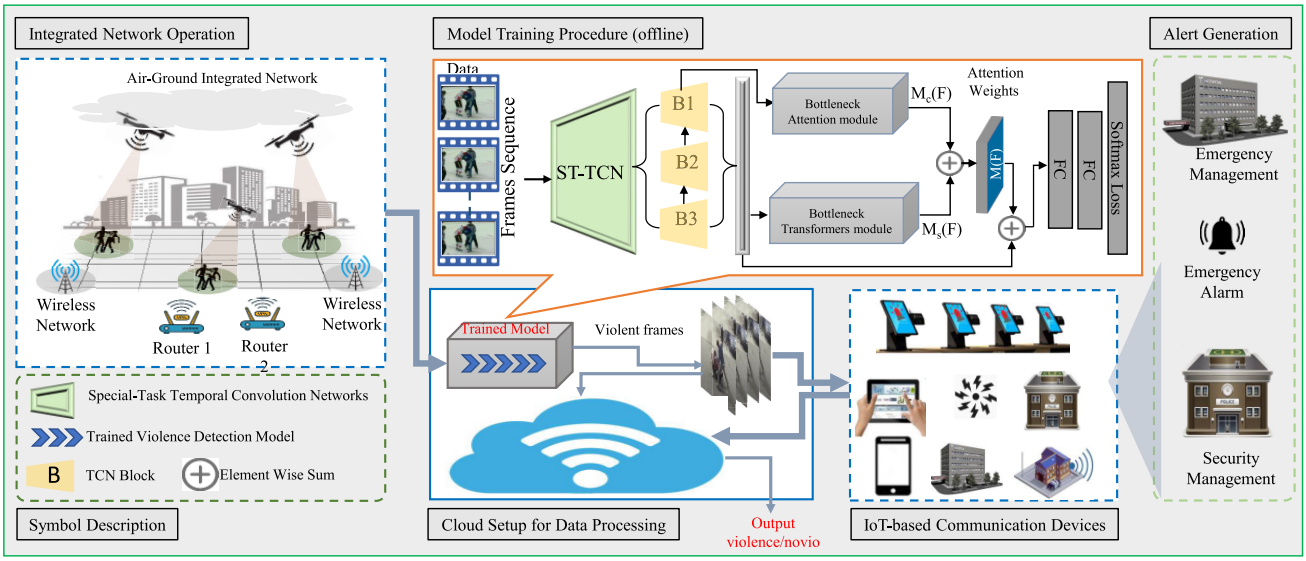

FIGURE 1. Overview of the proposed violence detection system with four fundamental modules: integrated Network Operation, model Training Procedure, cloud Setup for Data Processing, and IoT-based Communication Devices.

IIoT-based VD-Net to cover indoor and outdoor activities with low latency in real-time. The basic steps of the proposed framework are listed below. The first step demonstrates how to train the VD-Net for violent action detection offline. It deals with data acquisition using vision sensors with limited resources. At the same time, the second stage involves a screening process to collect critical information, such as identifying people or suspicious activities on the scene. Suppose the subjects and actions are identified in the second phase as violent or suspicious. In this case, if there are any violent frames, an alert is generated, and they are sent to the next step for a thorough investigation before the final violence detection phase.

Furthermore, the input data set D is separated, S_{tr} training data, S_{ts} testing data, and S_{vl} validation data, as shown as a pseudo-code in the **algorithm 1**. After validation, the trained model D_{mt} is obtained. A sequence Se_{tr} of frames from S_{tr} is fed into ST-TCN to produce a feature map F_{C-m} that is forward propagated into attention mechanisms for the final classification, as the output of the trained model D_{mt}. Similarly, the algorithm also reveals the pseudo-code of our system, which operates in real-time for violence detection. The input frames are extracted from various sources, such as surveillance cameras and unmanned aerial vehicles (UAVs). They are initially processed as a sequence S_e for significant violence F_{vio}. The inputs are supplied into system, which conduct a detailed study of the final output to determine whether it is violent or not if the frames F_{vio} have suspicious information and actions or activities. The proposed IIoT system transmits information and receives F_{vio} concerning violence from essential humans F_h. **Algorithm. 1**, represents the pseudo-code of the designed system.

Algorithm 1 Pseudo-Code of the Proposed IIoT-Based Violence Detection System

INPUTS: $\mathbf{V_{seq}} \leftarrow$ input sequence
Parameter Initialization:
1. $ST_{\mathbf{i}} \leftarrow$ Start model initialization
2. $D_{\mathbf{mt}} \leftarrow$ Call (Trained VD Model)
3. $S_{\mathbf{e}} \leftarrow$ recall (Input Sequence)

Recalling Main procedure for patterns:
1. **While** ($V_{\mathbf{seq}}$)
2. $S_{\mathbf{e}} \leftarrow$ Read (V_{seq})
3. $F_{\mathbf{vio}} \leftarrow ST_i(S_e)$
4. **if** $F_{\mathbf{vio}} == F_{(Fight,Shot)}$
5. initialize alert alarm and send details)
(recall IIoT Model) $D_{\mathbf{mt}}F_{(Fight,Shot)}$
6. **else** (if no violence detecting)
7. Continue and repeat step 2....
8. **end if**
10. **Continue surveillance**
11. **end while**

Return: Violence, nonviolence

A. SPATIAL TASK TEMPORAL CONVOLUTION NETWORK (ST-TCN)

We propose the ST-TCN blocks as a feature extractor as an alternative of traditional RNNs, which use feedback loops to propagate information through time. But our TCN uses temporal convolutions to capture long-term dependencies in the input sequence as illustrated in **Fig. 2(a)**. The input sequence is fed into a stack of convolution layers. In contrast, each layer applies a filter to a sequence of input values, with the other attributes determining the size of the receptive

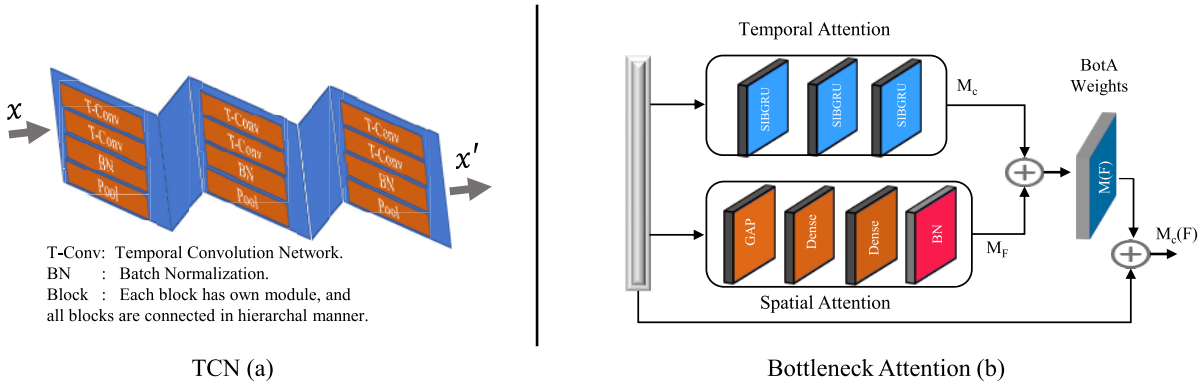

FIGURE 2. Overview of the proposed ST-TCN (a) and bottleneck attention (b) modules. All blocks are hierarchically interconnected in TCN and parallel in Bottleneck attention.

field. Therefore, by stacking multiple convolution layers with increased receptive field, TCN can capture dependencies over increasingly long periods. We designed three blocks of the ST-TCN in a hierarchical manner to learn features efficiently and make them parallelized for long sequence processing. Each block is connected hierarchically with input to capture deeper and hidden features from long sequences.

Each ST-TCN block consists of Input, a sequence of data points, and a tensor of shape (sequence length, input dimension), convolutional layers. The convolution filters have a fixed size and are convolved over the input sequence with a specified receptive field. After each layer, an activation function is applied to the layer's output with residual connections to improve the flow of gradient in training and alleviate the vanishing gradient problem. Our network often includes residual connections that bypass some of the convolutional layers. This allows the network to learn the input sequence's short- and long-term dependencies. Similarly, downsampling and upsampling are used to reduce the dimensionality of the output and extract the key features from the sequence. Through this, we capture long-term dependencies in the input sequence using a stack of layers, enhancing the system efficiency and scalability for real-time.

B. BOTTLENECK TRANSFORMER NETWORK (BTNET)

The bottleneck transformer architecture was introduced in [33] as a newer version of a standard transformer. The self-attention mechanism is applied to all the input tokens, which can be computationally expensive for large input sequences. The visual workflow of the modified BTNet and baseline is shown in **Fig. 3**. In a bottleneck transformer, a subset of input tokens is randomly selected and processed by a smaller number of attention layers before being combined with the remaining tokens and passed to the next layer. This reduces the number of attention layers needed, resulting in a more computationally efficient architecture. We used the same strategy, fine-tuned the traditional bottleneck transformer network (BTNet) for violence detection tasks, and trained it similarly to standard transformers using backpropagation and gradient descent optimization. We encoded the content position according to the height and width of the impute tensor and parallel connected with the original content to focus on salient cues. The proposed BTNet performed better than the standard while requiring fewer computational resources.

The learning strategy of the proposed BTNet with the input sequence is split into two parts: ''core'' and ''context'' sequence. The context sequence is processed separately by a set of attention layers. The core and context sequences are combined and fed to the forward layer. The combination is performed by simply concatenating the two sequences and passing them through a linear projection layer. The process can be repeated across multiple layers. While reducing the number of attention layers applied to the core sequence. Due to this strategy, the proposed BTNet achieves more computationally efficient results without losing performance. Additionally, we employed different forms of pattern dropout, which helps to regularize the network and prevent overfitting. This involves randomly dropping out entire patterns of attention weights during training rather than individual weights. The strategy encourages the network to learn more robust representations and improves its generalization ability for new data.

C. BOTTLENECK ATTENTION

The bottleneck attention mechanism [34] is a technique used in deep learning to improve the efficiency of attention-based models. Attention mechanisms are used to selectively focus on essential parts of input data when dealing with long sequences. Bottleneck attention reduces input dimensionality, cutting costs while preserving crucial information. Our proposed bottleneck includes a sequential and spatial layer, with an attention mechanism operating on the reduced-dimensional representation using standard dot-product attention. Our suggested module calculates the final

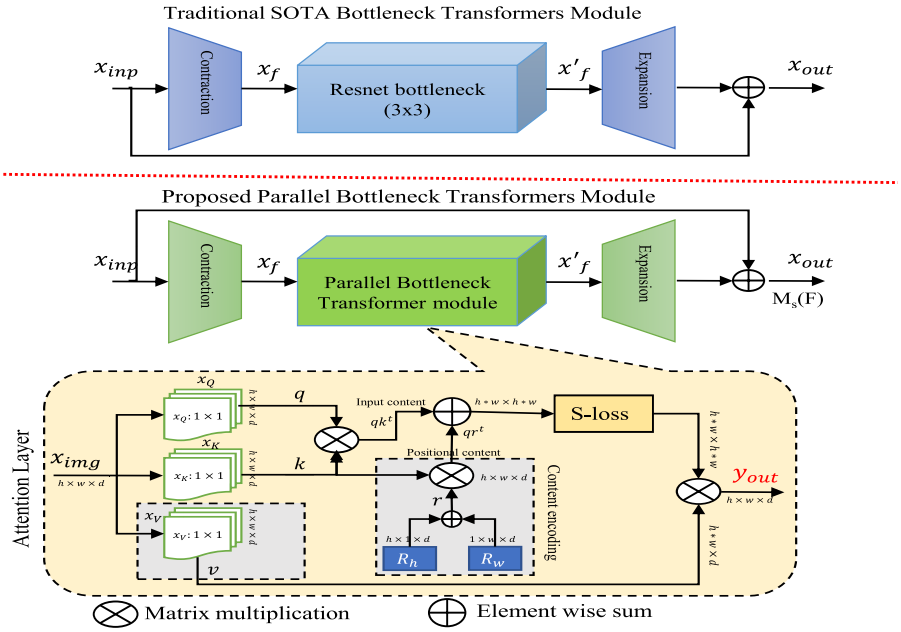

FIGURE 3. Top: traditional bottleneck transformer architecture. Bottom: the proposed bottleneck parallel transformer (BTNet).

feature map by:

$$F' = F + F * M(F) \tag{1}$$

$$M(F) = \alpha M_s(F) + M_c(F)) \tag{2}$$

$$\begin{aligned} M_s &= BN(FC(AP(F))) \\ &= BN(w_1(w_0 AP(F) + b_0) + b_1) \end{aligned} \tag{3}$$

$$M_c(F) = Seq(\int_{l1}^{n}, (\int_{l2}^{n}), (\int_{ln}^{n}(F))) \tag{4}$$

$$F' = (1 + M(F))F(x, \theta) \tag{5}$$

$$\frac{\alpha M(F)F(x,\theta))}{\alpha\theta} = M(F)\frac{\alpha F(x,\theta)}{\alpha\theta} \tag{6}$$

The notation $M_s(F)$ represents spatial attention features, $M_c(F)$ indicates the temporal attention cues, and $M(F)$ shows the final attention weights. The procedure outlined above demonstrates the direct correlation between gradients and attentions value. Higher attention values require higher gradient values, and vice versa. The symbol θ denotes the constraints employed in extraction of features. As a result, we use the attention process like channel-wise to observe, detect, and selectively concentrate on salient cues. Instead of choosing an object region distinct from channel attention, spatial attention removes clutter and chooses important spatial places via dense layers. Due to the use of this complementary information, combining these two attentional systems is essential for tasks involving violence.

Our implementation of attention maps has resulted in a highly effective approach, allowing us to concentrate on each branch's specific goals in the input tensor with precision and accuracy. Our attention module consists of an input sequence, a layer that reduces its dimensionality, such as a linear projection. This layer is often called the "bottleneck layer" because it acts as a bottleneck that reduces the amount of information passed to the attention mechanism. Once the input sequence is transformed into a lower-dimensional representation, it is fed into an attention mechanism that evaluates the importance of each element in the sequence by assigning it a weight proportional to its relative significance. The weighted sequence and the resulting attention weights are used to weight the input sequence, giving more weight to important elements and less weight to unimportant elements. This produces a weighted sequence that emphasizes the most relevant information. The resulting attention weights are used for the input sequence and then fed into the rest of the model. **Fig. 2 (b)**, illustrates the visual representation of the proposed bottleneck attention module.

IV. EXPERIMENTAL SETUP AND RESULTS

This section presented the experimental evaluations, comparison, and discussion of the proposed VD-Net system. We also summarize the dataset information, hardware configuration, and qualitative and quantitative results. Our proposed VD-Net is evaluated via various metrics, including the receiver operating characteristic (ROC) curve, F-measurements, precision/recall, and confusion matrix. More detailed information on these evaluations is presented in subsequent sections.

A. DATASET

The rapid advancement of technology has led various sectors to actively engage in violence detection, aiming to address data challenges in surveillance for safety and security

FIGURE 4. Visually representation of the utilized datasets. (a) show the surveillance, (b) show the hockey, (c) show the movie, and (d) show the violent dataset.

TABLE 1. Violence databases used to evaluate the proposed VD-Net: A statistical overview of the utilized datasets.

Database Name	Totl samples	Frame Rate (fps)	resolution (pixel)	Length (Sec)
Hockey Fight	1000	25	360x240	1.6
Violent Flow	246	20–30	varies	5
Surveil Fight	300	20–30	varies	3-5
Movie Fight	200	25	360x240	1-1.9s

purposes. One key challenge is the need to limit surveillance data to specific indoor and outdoor activities. To overcome these limitations, there is a growing interest in leveraging technology to enhance surveillance capabilities by utilizing different datasets to generalize the model's capability. In this study, four datasets, including surveillance fight [35], violent flow [36], hockey, and movie fight [37], are employed to develop a violence detection model. Each dataset is divided into violent and nonviolent classes and split into training, validation, and testing sets following standard procedures. The training set comprises 70 % of the data, while the validation and testing sets account for 20 % and 10 %, respectively.

The surveillance fight dataset [35] includes videos of violence captured by surveillance cameras in various locations that are mostly used to develop and evaluate algorithms for detecting violent behavior in real-time. Likewise, a violent flow [36] includes videos shot in factories, offices, and other settings indoors, outdoors, during the day, and at night. This dataset consists of unaltered surveillance videos. Moreover, the National Hockey League released the hockey and movie fight datasets [37]. The initial movie dataset comprises 200 video clips featuring fight scenes from action movies, while the non-fight videos were sourced from public action datasets. The hockey dataset includes 1000 clips and 500 for each fight and non-fight. In contrast to the hockey dataset, where all sequences were recorded in the same format and size, the movie dataset used various resolutions and formats. Still, it was more homogenous in content and format. **Fig. 4** visually represents a few samples from each dataset, while **Table 1** provides statistical details of each dataset.

B. SYSTEM CONFIGURATION

The proposed architecture uses a Jetson device as an edge server to gather data streams from devices connected to an IoT network. For inference tasks, the Jetson AGX Orin 64GB module is specifically used on the edge device for screening purposes. It incorporates the advanced NVIDIA Orin system-on-chip (SoC), seamlessly integrating multiple ARM cores, cutting-edge GPU architecture, and dedicated AI accelerators. The AGX Orin has a generous 64GB of onboard memory and ample storage capacity for AI models and efficient data processing. Moreover, it offers extensive connectivity options, including Ethernet, PCIe, USB, and MIPI CSI interfaces, ensuring effortless integration with a wide range of sensors and peripherals.

The VD-Net is implemented using the widely-used deep learning framework TensorFlow (version 2+), and Adam is used as the optimizer. Similarly, the training platform uses GeForce 3080-Ti RTX GPUs. The model is trained with the early stopping technique, a defined 16 batch size, with

TABLE 2. Ablation study of the designed violence detection network with baseline using Violence databases, where the B indicates the baseline and MB indicates the modified baseline module.

Method	Hockey Fight (%)	Move Fight (%)	Surveillance Fight (%)	Violent Flow (%)
Baseline (B)	95.01	91.85	77.95	88.45
B + TCNs	96.20	93.50	83.00	92.00
B + ST-TCN	97.45	95.70	88.56	93.97
MB + ST-TCN	**98.50**	**99.00**	**92.50**	**97.00**

TABLE 3. Ablation study of the designed violence detection network using hockey dataset with different sequence lengths (SL). The B represents the baseline, and MB represents the modified version of the baseline model.

Method	SL	Result (%)	SL	Result (%)	SL	Result (%)	SL	Result (%)
Baseline (B)	-	80.70	-	84.45	-	85.75	-	81.0
B + TCNs	10	82.00	15	86.21	20	92.11	25	82.35
B + ST-TCN	-	84.80	-	88.97	-	94.45	-	84.44
MB + ST-TCN	**-**	**85.16**	**-**	**90.50**	**-**	**98.01**	**-**	**87.95**

other supporting functions that expedite the model training and performance. The data were split via the standard procedure, a 70:20:10 percent ratio for training, validation, and testing.

C. ABLATION STUDY

Initially, as an ablation study, we used four publicly available datasets of violent behaviors and evaluated the accuracy of the baseline model, called bottleneck transformers [29], and further evaluated the model by adding temporal convolution networks (TCNs) [38] (Baseline + TCNs). Consider these as a baseline and propose a lightweight VD-Net model using modified bottleneck transformers (Mbaseline) with spatial task TCN (MBaseline + ST-TCN). The quantitative experimental evaluation of the proposed VD-net and the baseline method is illustrated in **Table 2**.

The proposed technique achieved 98.50 % accuracy on the hockey fight, 97.00 % accuracy on the violent flow, 92.50 % accuracy on the surveillance camera fight, and 99.00 % accuracy on the movie fight dataset over the baseline model. Our method improved the violence recognition rate over the baseline model with reduced computational cost as mentioned in **Tabel 6**. While the accuracy of the baseline model using TCNs to add the ST-TCN blocks improved, the TCNs' performance. The experiments demonstrated that the modified ST-TCN and bottleneck transformers module significantly enhance the baseline's ability to extract spatiotemporal information. The bottleneck connection made the proposed model more sensitive to extracting the salient information of violent behaviors by cross-channel interaction and increased the precision of the model. Furthermore, we conducted an extensive ablation study on various frame sequences to choose an appropriate sequence to better recognize violent/nonviolent actions. The hockey fight dataset is used for sequence selection as illustrated in **Table 3**.

D. RESULTS AND EVALUATIONS

This section provides a comprehensive evaluation of the proposed VD-Net system based on the testing results obtained from each dataset. We trained and evaluated conventional bottleneck transformers and TCNs for each dataset to compare with the proposed VD-Net as shown in **Tabel 2** to highlight the strengths and weaknesses of each method, enabling a better understanding of the proposed approach.

The modern industrialized world and smart cities have established surveillance systems that enable activity tracking on a layer above to combat and monitor violent things. Despite the fact that most current approaches use non-surveillance datasets, automating this system presents quite a few challenges, as mentioned earlier. To support this technology, we primarily concentrate on the VD-based surveillance setup and assess our method's effectiveness using newly published datasets [35], [36] and compare with SoTA. The surveillance camera dataset endorsed serves as a baseline for security surveillance in industrial and commercial areas. Furthermore, we checked our system on non-surveillance dataset such as hockey and movie fights [37] for more generalization.

We thoroughly review SoTA capabilities and compare them to the proposed VD-Net. We test the proposed system using four datasets, and the results are presented in **Table 4** with the confusion matrix depicted in **Fig. 5**. The proposed system achieves high precision in detecting violence in both indoor and outdoor datasets. For the convenience of readers and researchers, we produced the ROC (receiver operating characteristic) curves with accuracy in **Fig. 6** to display the suggested model performance on surveillance and non-surveillance datasets, respectively. It is common for VD techniques to define any progressing activity as violent in sports, where athletes collide or hit one another, e.g., in a hockey fight. As a result, one method of detecting aggression

TABLE 4. The thorough assessment outcomes of the suggested VD-Net are presented based on precision, accuracy, F1-score, and recall utilizing public benchmarks.

	Violent		Non-Violent			Measurement Matrices		
Dataset	Recall	Precision	Recall	Precision	F1 score	w_acc	unw_acc	Accuracy (%)
Hockey Fight	0.98	0.98	0.98	0.99	0.98	0.98	0.98	98.50
Violent Flow	0.94	1.00	0.99	0.95	0.96	0.96	0.97	97.00
Surveillance Fight	0.99	0.85	0.90	0.97	0.92	0.94	0.92	92.50
Movie Fight	0.98	1.00	0.99	0.99	0.98	1.00	0.98	99.00

TABLE 5. We conducted a comparative analysis across the proposed and state-of-the-art (SoTA) methods using benchmark datasets in terms of accuracy (%) and their learning strategies.

Architecture	Hockey Fight(%)	Movie Fight(%)	Violent Flow(%)	Surveillance Camera(%)
ResNet-50 [39]	95.50	91.00	93.87	85.21
I3D [40]	94.50	87.80	88.89	84.67
Adoptive Resolution Network [41]	92.11	90.01	94.37	77.34
Temporal Shift module [42]	97.50	88.70	93.95	79.01
Temporal Excitation Network [43]	90.70	90.00	92.93	82.50
MiNet-3D [44]	94.71	100.00	91.41	–
ViF [45]	82.90	–	85.00	–
OViF [36]	87.50	–	88.00	–
ViF + OViF [36]	86.30	–	–	–
MoWLD-BoW [37]	90.90	89.50	–	–
3D-CNNs [46]	96.00	90.20	98.00	–
Two-casecade TSM [47]	98.05	–	96.93	–
SSHA [48]	97.05	–	97.90	–
FightCNN-based Attention [49]	95.00	–	–	71.00
ViT Large-16 [50]	98.00	99.50	97.00	84.60
RCNN-based Darknet [32]	98.00	–	–	74.00
3D CNN-based VDNet [51]	–	–	94.00	88.70
Proposed VD-Net	**98.50**	**99.00**	**97.00**	**92.50**

is to watch how players approach one another. However, there is a real risk viewers will mistake a player's hug during a winning celebration for a violent gesture. However, our suggested system encoding special and temporal information is employed to avoid these mistakes and differentiate violent frames. Overall our system is convenient and generalized for indoor and outdoor actions to easily predict/differentiate violent and nonviolent actions by their testing outcomes.

E. DISCUSSION AND COMPARISON WITH SOTA

The detail discussion and analysis of the our system illustrated in this section. Our proposed model leverages a hierarchical integration of bottleneck layers and a specialized temporal convolutional network to achieve superior violence recognition/detection results. The comparison is presented and demonstrated at **Table 5**, highlighting the efficiency of our designed system.

The proposed AI-based method achieves a 97 % recognition rate on violent flow, outperforming most SoTA algorithms. The proposed method has just 0.351 % lower accuracy than ViT large [52] on the movie fight dataset but higher on a hockey fight, and violence flows as well as the computation complexity of our model is lower than ViT large [52], which is shown in the subsequent section. Our recognition rate on all four utilized datasets was marginally higher than the pre-trained ResNet50 [39], I3D [40], AR-Net [41], temporal shift module, and temporal adaptation encoder [42], [43]. All these pre-trained weights are used and trained on violent datasets, and their results are reported in **Table 5** for comparative analysis.

The proposed model produces the best accuracy as shown in **Table 5**. The authors claimed the lighting of a model in articles for violence detection by combining different modules. The authors used depth-wise separable convolutions and a bottleneck learning strategy to enhance the Vd-Net performance. Furthermore, our method hierarchically used ST-TCN blocks as a primary component of learning the spatiotemporal cues, as seen in the main framework. The accuracy does not significantly increase in

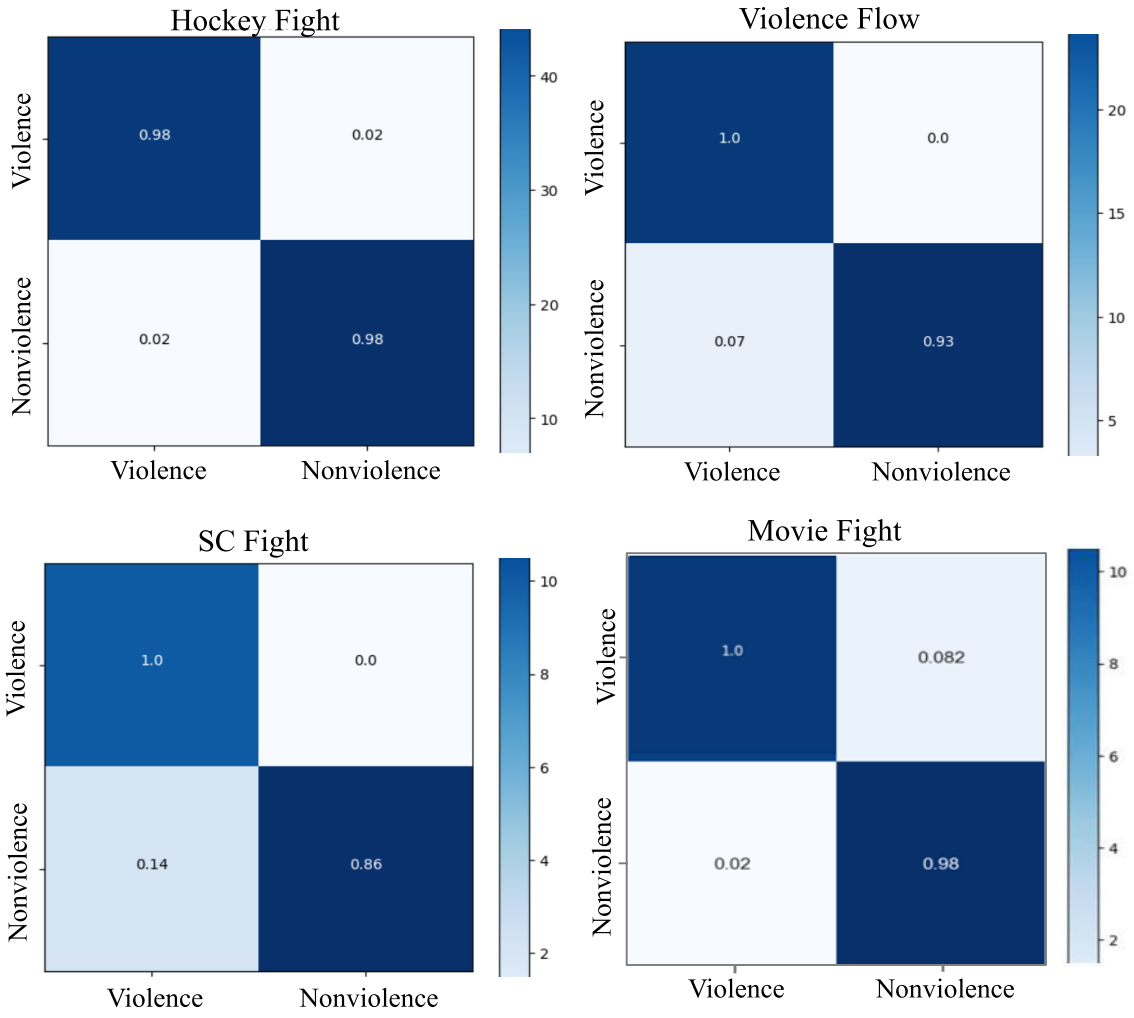

FIGURE 5. The confusion among each dataset's actual and predicted values to better understand the model's performance.

some cases, but our model has fewer parameters and requires less computation, which can easily be implemented on edge devices. The computational complexity of the model is the main objective of this research to be installed on edge devices in an IoT environment, which is explained in the subsequent section.

1) COMPLEXITY ANALYSIS

We conducted some experiments about computations of the proposed VD-Net algorithms, which clearly illustrates the significant benefits of our approach as shown in **Table 6**. We compared the computational costs of the proposed VD-Net with baselines in terms of parameters, model size, and FLOPs (Floating-point Operations). For instance, we utilized a torch model summary to determine this information representing the model's computational complexity variables without requiring a manual calculation. Additionally, TensorBoard is used to record pertinent data, which includes the total time spent during training and testing for all epochs. Our suggested methodology significantly reduces the number of parameters utilized for model training compared to SoTA algorithms. In addition, the training time and model size have significantly decreased, creating ideal circumstances for implementing IoT-based edge devices.

TABLE 6. Computational complexity of the proposed VD-Net with baselines.

Algorithm	Size (MB)	Parameters (MB)	GFLOPs
3D-CNN [46]	2647.70	297.56	–
TSM [42]	397.71	89.93	98
I3D [40]	1000.20	146.88	–
TEA [43]	479.78	91.95	70
AR-Net [41]	–	–	33.47
P-TSM [53]	369.93	89.85	–
Our VD-Net	**188.00**	**49.28**	**15.30**

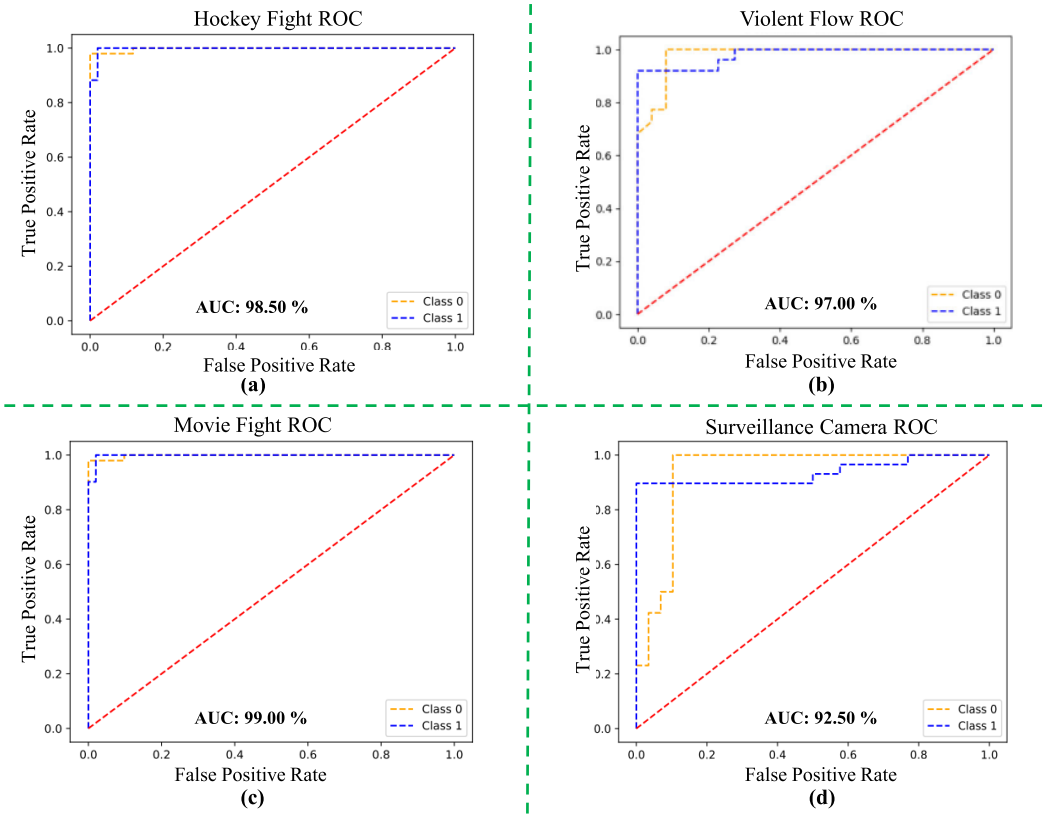

FIGURE 6. The proposed violence detector ROC carves.

V. CONCLUSION

Technology and surveillance infrastructure have advanced to improve security and protect assets to ensure public safety. Manual monitoring is tedious and time-consuming, especially when analyzing potential threats or violence. To overcome this challenge, several existing methods have attempted to address the issue by utilizing distinct algorithms, which process frames locally without utilizing IoT settings. However, these methods still have limitations that must be overcome to achieve optimal performance and efficiency in surveillance and security systems.

VD-Net offers an innovative and integrated approach to surveillance, providing more efficient and effective security for individuals and organizations by addressing the limitations of manual methods and IoT integration. The VD-Net analyzes input frames, extracting crucial information such as humans and violent objects, and shares relevant data within the IIoT network to make a final decision and alert concerned parties in case of violent incidents. Through evaluation, our method achieved significant improvements in accuracy compared to existing methods in the literature. These results highlight the suitability of our approach for deployment in security systems over edge devices and present a significant step in enhancing security and safety in various settings.

Moreover, we plan to further improve the proposed VD-Net framework by exploring real-time data processing techniques and edge computing to reduce processing delays and enable faster detection of violent scenes. To achieve this, we may consider deploying our VD-Net model on more powerful edge devices or cloud servers with GPUs to enable more complex computations in real time.

To further validate the effectiveness and suitability of our approach for real-world deployment, we plan to test our method on more diverse datasets with a broader range of violent and nonviolent activities. This will allow us to assess the model's performance under different scenarios and conditions, identify potential limitations, and make necessary adjustments. Overall, our proposed approach has significant potential for enhancing security and safety in various settings, and we aim to continue improving and refining it to achieve optimal performance and efficiency.

ACKNOWLEDGMENT

The authors acknowledge the invaluable contribution of artificial intelligence (AI) tools in enhancing the efficiency and advancements in their research endeavors.

REFERENCES

[1] U. Dikwatta and T. Fernando, "Violence detection in social media—Review," *Vidyodaya J. Sci.*, vol. 22, no. 2, pp. 7–16, 2019.

[2] Z. Shao, J. Cai, and Z. Wang, "Smart monitoring cameras driven intelligent processing to big surveillance video data," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 105–116, Mar. 2018.

[3] W. So. (2019). *Perceived and Actual Leading Causes of Death Through Interpersonal Violence in South Korea As of 2018*. [Online]. Available: https://www.statistic.com/statistics/

[4] N. Mumtaz, N. Ejaz, S. Habib, S. M. Mohsin, P. Tiwari, S. S. Band, and N. Kumar, "An overview of violence detection techniques: Current challenges and future directions," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4641–4666, May 2023.

[5] R. Nawaratne, S. Kahawala, S. Nguyen, and D. De Silva, "A generative latent space approach for real-time road surveillance in smart cities," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4872–4881, Jul. 2021.

[6] M. Fu, S. Sun, Q. Liang, X. Tong, and Q. Liu, "Exciting-inhibition network for person reidentification in Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15059–15069, Jun. 2020.

[7] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," *IEEE Access*, vol. 7, pp. 107560–107575, 2019.

[8] T. Perperis, T. Giannakopoulos, A. Makris, D. I. Kosmopoulos, S. Tsekeridou, S. J. Perantonis, and S. Theodoridis, "Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies," *Expert Syst. Appl.*, vol. 10, p. 14, May 2011.

[9] M. S. Vural and M. Gök, "Criminal prediction using naive Bayes theory," *Neural Comput. Appl.*, vol. 28, no. 9, pp. 2581–2592, Sep. 2017.

[10] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0203668.

[11] V. Lam, S. Phan, D.-D. Le, D. A. Duong, and S. Satoh, "Evaluation of multiple features for violent scenes detection," *Multimedia Tools Appl.*, vol. 76, no. 5, pp. 7041–7065, Mar. 2017.

[12] T. Aremu, L. Zhiyuan, R. Alameeri, M. Khan, and A. E. Saddik, "SSIVD-Net: A novel salient super image classification & detection technique for weaponized violence," 2022, *arXiv:2207.12850*.

[13] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly supervised violence detection in surveillance video," *Sensors*, vol. 22, no. 12, p. 4502, Jun. 2022.

[14] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2945–2956, Dec. 2017.

[15] S. Ul Amin, M. Ullah, M. Sajjad, F. A. Cheikh, M. Hijji, A. Hijji, and K. Muhammad, "EADN: An efficient deep learning model for anomaly detection in videos," *Mathematics*, vol. 10, no. 9, p. 1555, May 2022.

[16] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.

[17] X. Zhou, Y. Chen, and Q. Zhang, "Trajectory analysis method based on video surveillance anomaly detection," in *Proc. China Autom. Congr. (CAC)*, Oct. 2021, pp. 1141–1145.

[18] M.-F. Guo, X.-D. Zeng, D.-Y. Chen, and N.-C. Yang, "Deep-learning-based earth fault detection using continuous wavelet transform and convolutional neural network in resonant grounding distribution systems," *IEEE Sensors J.*, vol. 18, no. 3, pp. 1291–1300, Feb. 2018.

[19] F. Barros, S. Aguiar, P. J. Sousa, A. Cachaço, P. J. Tavares, P. M. G. P. Moreira, D. Ranzal, N. Cardoso, N. Fernandes, R. Fernandes, R. Henriques, P. M. Cruz, and A. Cannizzaro, "Displacement monitoring of a pedestrian bridge using 3D digital image correlation," *Proc. Struct. Integrity*, vol. 37, pp. 880–887, Jan. 2022.

[20] M. Khan, W. Gueaieb, A. E. Saddik, G. De Masi, and F. Karray, "An efficient violence detection approach for smart cities surveillance system," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, Sep. 2023, pp. 1–5.

[21] S. Mishra, N. Srinivasan, and U. S. Tiwary, "An affective video database using multimedia content analysis rated on Indian samples," 2022, *arXiv:2210.09785*.

[22] M. Khan, M. Saad, A. Khan, W. Gueaieb, A. E. Saddik, G. De Masi, and F. Karray, "Action knowledge graph for violence detection using audiovisual features," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2024, pp. 1–5.

[23] G. Liu, Z. Wang, H. Zhang, X. Guo, Y. Wang, and C. Zhang, "A novel violent video detection method based on improved C3D and transfer learning," in *Proc. 3rd Int. Conf. Comput. Inf. Big Data Appl.*, Mar. 2022, pp. 1–7.

[24] E. Fenil, G. Manogaran, G. N. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. J. C. N. Ahilan, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Comput. Netw.*, vol. 151, pp. 191–200, Mar. 2019.

[25] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets," *Neurocomputing*, vol. 371, pp. 188–198, Jan. 2020.

[26] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," *Appl. Artif. Intell.*, vol. 34, no. 4, pp. 329–344, Mar. 2020.

[27] B. D. Mishra, L. Huang, N. Tandon, W.-T. Yih, and P. Clark, "Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension," 2018, *arXiv:1805.06975*.

[28] I. Tourni, L. Guo, T. H. Daryanto, F. Zhafransyah, E. E. Halim, M. Jalal, B. Chen, S. Lai, H. Hu, M. Betke, P. Ishwar, and D. T. Wijaya, "Detecting frames in news headlines and lead images in U.S. gun violence coverage," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 4037–4050.

[29] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16514–16524.

[30] G.-P. Ji, Y.-C. Chou, D.-P. Fan, G. Chen, H. Fu, D. Jha, and L. Shao, "Progressively normalized self-attention network for video polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 142–152.

[31] Q. Li, R. Yang, F. Xiao, B. Bhanu, and F. Zhang, "Attention-based anomaly detection in multi-view surveillance videos," *Knowl.-Based Syst.*, vol. 252, Sep. 2022, Art. no. 109348.

[32] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, A. Ullah, S. W. Baik, F. Cuzzolin, J. J. P. C. Rodrigues, and V. H. C. de Albuquerque, "An intelligent system for complex violence pattern analysis and detection," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10400–10422, Dec. 2022.

[33] A. Fan, E. Grave, and A. Joulin, "Reducing transformer depth on demand with structured dropout," 2019, *arXiv:1909.11556*.

[34] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "A simple and light-weight attention module for convolutional neural networks," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 783–798, Apr. 2020.

[35] S. Akti, G. A. Tataroglu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *Proc. 9th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2019, pp. 1–6.

[36] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented VIolent flows," *Image Vis. Comput.*, vols. 48–49, pp. 37–41, Apr. 2016.

[37] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2011, pp. 332–339.

[38] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[41] Y. Meng, C. C. Lin, R. Panda, P. Sattigeri, L. Karlinsky, A. Oliva, K. Saenko, and R. Feris, "AR-Net: Adaptive frame resolution for efficient action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 86–104.

[42] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092.

[43] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 906–915.

[44] W. Wang, S. Dong, K. Zou, and W. Li, "A lightweight network for violence detection," in *Proc. 5th Int. Conf. Image Graph. Process. (ICIGP)*, Jan. 2022, pp. 15–21.

[45] 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.

[46] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2472, May 2019.

[47] Q. Liang, Y. Li, B. Chen, and K. Yang, "Violence behavior recognition of two-cascade temporal shift module with attention mechanism," *J. Electron. Imag.*, vol. 30, no. 4, p. 43, Jul. 2021.
[48] Y. Zhang, Y. Li, and S. Guo, "Lightweight mobile network for real-time violence recognition," *PLoS ONE*, vol. 17, no. 10, Oct. 2022, Art. no. e0276939.
[49] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, Mar. 2022.
[50] L. Ciampi, C. Santiago, J. P. Costeira, F. Falchi, C. Gennaro, and G. Amato, "Unsupervised domain adaptation for video violence detection in the wild," in *Proc. IMPROVE*, 2023, pp. 37–46.
[51] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3D convolutional neural networks," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.
[52] S. Akti, F. Ofli, M. Imran, and H. K. Ekenel, "Fight detection from still images in the wild," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 550–559.
[53] Y. Zhang, Y. Li, S. Guo, and Q. Liang, "Not all temporal shift modules are profitable," *J. Electron. Imag.*, vol. 31, no. 4, Jul. 2022, Art. no. 043030.

MUSTAQEEM KHAN (Member, IEEE) received the Ph.D. degree in software engineering from Sejong University, Seoul, Republic of Korea. Currently, he is a Lead Researcher with MCR Laboratory, Mohamed Bin Zayed University of Artificial Intellegence (MBZUAI). His primary research focus encompasses affective computing, computer vision, and emotion recognition. Additionally, his academic pursuits extend to areas such as audio digital signal processing, speech processing, speech synthesis, image and video processing, energy analytics, consumption predictions, and generation. With a notable presence in the academic community, he holds roles as an Associate Editor and a Guest Editor, along with serving as a professional reviewer for esteemed journals and conferences. He is a Gold Medalist.

ABDULMOTALEB EL SADDIK (Fellow, IEEE) is an Enterprise Professor with MBZUAI, United Arab Emirates, while holding the esteemed position of a Distinguished University Professor with the University of Ottawa, Canada. He is widely regarded as an internationally recognized scholar, and significantly advanced the fields of intelligent multimedia computing, communications, and applications. His research endeavors are focused on leveraging AI, the IoT, SN, AR/VR, haptics, and 5G technologies to establish digital twins that enhance citizens' quality of life. Through his work, individuals can engage in real-time interactions with one another and their smart digital representations in the metaverse, ensuring security, and fostering seamless connectivity. With a prolific academic career, he has coauthored ten books and published over 800 research contributions. His exemplary track record includes securing research grants and contracts exceeding 20 million. Notably, he has authored the influential book *Haptics Technologies: Bringing Touch to Multimedia*. He was recognized for his outstanding contributions, and elected as a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. Moreover, he has chaired more than 50 conferences and workshops and mentored over 150 researchers. In addition, he holds the title of ACM Distinguished Scientist. He currently serves as the Editor-in-Chief for the *ACM Transactions on Multimedia Computing, Communications, and Applications* (ACM TOMM).

WAIL GUEAIEB (Senior Member, IEEE) received the bachelor's and master's degrees in computer engineering and information science from Bilkent University, Turkey, in 1995 and 1997, respectively, and the Ph.D. degree in systems design engineering from the University of Waterloo, Canada, in 2001. He is currently a Professor with the School of Electrical Engineering and Computer Science (EECS), University of Ottawa, Canada. He also founded and directed the Machine Intelligence, Robotics, and Mechatronics (MIRaM) Laboratory, EECS. He is the author/coauthor of over 120 patents and articles in highly reputed journals and conferences. His research interests span the fields of intelligent mechatronics, robotics, and applied computational intelligence. He is currently an Associate Editor of the *ASME Journal of Dynamic Systems, Measurement, and Control*; and the *International Journal of Robotics and Automation*. He has served as an Associate Editor, a Guest Editor, and the Program (Co-)Chair for several international journals and conferences, such as IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, IEEE/ASME TRANSACTIONS ON MECHATRONICS, and the IEEE Conference on Decision and Control.

GIULIA DE MASI (Senior Member, IEEE) received the master's degree (magna cum laude) from University L'Aquila and the Ph.D. degree from the University of Rome La Sapienza, in collaboration with the Max Planck Institute, Dresden, and City, University of London. She is currently a Principal Scientist with the Autonomous and Robotic Research Center, Technology Innovation Institute, Abu Dhabi, United Arab Emirates. She was a Postdoctoral Researcher with the Polytechnic University of Marche; and a Visiting Researcher with the Hitachi Research Laboratory, Nara, Japan. In 2008, she started working in the research and development field with the Snamprogetti Center of Excellence, Italy, and joined the Department of Advanced Engineering Services and Technology Innovation Projects. She was with several academic institutions in United Arab Emirates, before joining the Technology Innovation Institute, as a Principal Scientist. She has two patents, more than 60 peer-reviewed publications (including journals and conferences), and more than 40 reports for the industry. Her main fields of expertise are collective intelligence, machine learning, deep learning, and optimization, with applications to multi-robot systems. She has been nominated as a Women in Data Science (WiDS) Ambassador of United Arab Emirates. She has also been recently awarded as the Women in Engineering (WIE) Propel Laureate IEEE by Oceanic Engineering Society in liaison with WIE.

FAKHRI KARRAY (Fellow, IEEE) is a renowned academic and researcher in the field of artificial intelligence. He is currently a Professor and a Provost with the Mohamed Bin Zayed University of Artificial Intelligence, United Arab Emirates, and holds prestigious positions with the University of Waterloo, Canada. He has made significant contributions to the field, authoring numerous publications, co-inventing patents, and serving in leadership roles at international conferences and journals. His research focuses on intelligent systems, operational AI, smart devices, and human–machine interaction through speech, gesture, and natural language processing. He is a fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada, reflecting his exceptional achievements and impact in the field of artificial intelligence.