جامعة ابن طفيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵉⴱⵏ ⵟⴼⵉⵍ
Ibn Tofaïl University
Faculté des Sciences

Université Ibn Tofail
Faculté des Sciences, Kénitra

Mémoire de Projet de Fin d'Etudes

Master Intelligence Artificielle et Réalité Virtuelle

Advanced Dual-Stream Framework for Violence Detection

Établissement d'accueil :

Laboratoire Systèmes Eléctroniques, Traitement de l'Information et Intelligence Artificielle, Mécanique et Energétique (SETIME)

Elaboré par : Mlle Khadija ETTOUIL

Encadré par : Mr Tarik BOUJIHA (ENSA KENITRA (UIT))

Mr Mohamed KAS (UTBM)

Soutenu le 20/09/2024, devant le jury composé de :

- Mme Raja TOUAHNI (FS Kenitra (UIT))
- Mr Rochdi MESSOUSSI (FS Kenitra (UIT))
- Mr Anass NOURI (FS Kenitra (UIT))
- Mr Tarik BOUJIHA (ENSA KENITRA (UIT))
- Mme Fadoua GHANIMI (FS Kenitra (UIT))
- Mr Idriss MOUMEN (FS Kenitra (UIT))

Table des matières

Acknowledgements

First and foremost, I am profoundly grateful to God, whose blessings and guidance have been the cornerstone of all my achievements. His grace has given me the strength, perseverance, and clarity to complete this work.

I would like to extend my heartfelt thanks to my mother, my father, and my entire family. Your unwavering love, patience, and encouragement have been my greatest source of strength throughout this journey. I am deeply appreciative of the sacrifices you have made and the support you have provided every step of the way.

I also wish to express my sincere appreciation to Pr. Raja TOUAHNI, the coordinator of my master's program. Your tireless commitment and exceptional mentorship have been instrumental in my academic success. I am truly grateful for the wisdom and guidance you have offered.

Additionally, I extend my deep thanks to the academic staff and professors, whose dedication and expertise have enriched my learning experience and contributed greatly to my personal and professional growth.

Abstract

With the increasing installation of surveillance cameras in various locations to ensure public safety, security, and asset protection, the demand for intelligent video surveillance has grown significantly. Violence detection, a key application of intelligent surveillance, plays a vital role in public safety, behavior monitoring, and law enforcement. Detecting violent events or behaviors in video sequences has been the focus of various research efforts over the years, resulting in a wide range of techniques and features being developed. This report presents a real-time violence detection framework that leverages a two-stream approach, incorporating both RGB and pose estimation data to address these challenges.

Our method utilizes NASNetMobile for feature extraction, LSTM for temporal modeling, and an attention mechanism to selectively focus on the most relevant features. By integrating both visual appearance and pose information, the model effectively captures and interprets complex human movements. This dual-stream architecture, enhanced by an attention mechanism, enables the model to adaptively focus on either RGB or pose data depending on the context, making it more robust across diverse scenarios.

To validate this approach, extensive experiments were conducted across multiple video-level benchmarks. In line with prior research, we emphasize the importance of ongoing analysis and comparison of state-of-the-art approaches. Unlike typical methods that often assess models on the same dataset, cross-dataset evaluations were considered, highlighting the limitations in model generalization. The experiments confirmed the effectiveness of combining RGB data, pose estimation, and attention mechanisms for improving violence detection performance, surpassing existing methods.

This report also discusses the challenges researchers face in this domain, such as achieving robust generalization across real-world scenarios. The categorization of violence detection techniques—handcrafted feature-based, deep learning, and hybrid approaches—has been critically analyzed, addressing the strengths and weaknesses of each. Additionally, research gaps and future directions are outlined, contributing to the ongoing effort in advancing violence detection technology and laying the foundation for further developments in the field.

Keywords :

- Intelligent surveillance
- Violence detection
- Computer vision
- RGB Data
- Pose Estimation
- NASNetMobile
- LSTM
- Attention Mechanism
- Video Analysis
- Behavior Monitoring

Introduction

The rapid proliferation of video surveillance systems across various sectors, from public safety to private security, has created an urgent need for advanced technologies that can automatically detect and respond to violent behavior in real-time. Traditional methods for violence detection have largely relied on manual monitoring or simplistic algorithms that struggle to balance accuracy and computational efficiency. As a result, these methods often fall short when applied to complex and dynamic real-world environments where violence can manifest in diverse and unpredictable ways.

In recent years, the field of computer vision has made significant strides in improving the accuracy of automated violence detection systems. However, many existing approaches still face substantial challenges. Key among these is the difficulty of effectively capturing and interpreting the nuanced movements and interactions that characterize violent behavior. Simple appearance-based models may overlook critical motion details, while purely motion-based methods might miss the contextual cues provided by the visual appearance of a scene. Additionally, the high computational cost associated with advanced deep learning models often limits their applicability in real-time scenarios, particularly on resource-constrained devices like mobile phones or edge computing platforms.

To address these challenges, this report presents a novel two-stream violence detection framework that integrates both RGB (color image) data and pose estimation data to deliver a more comprehensive analysis of video content. By leveraging NASNetMobile, a state-of-the-art convolutional neural network (CNN) for feature extraction, and LSTM (Long Short-Term Memory) networks for temporal modeling, this framework is designed to capture both the spatial and temporal aspects of violent actions. The RGB stream provides detailed visual information, while the pose estimation stream focuses on human body movements, allowing the model to analyze and interpret complex behaviors more accurately. The inclusion of an attention mechanism further enhances the model's capability by dynamically focusing on the most relevant features in each stream, improving detection accuracy and robustness across various scenarios.

The proposed framework has been subjected to rigorous testing across multiple benchmark datasets, each representing different forms of violence in varied settings. Notably, our method achieved state-of-the-art performance on the Hockey-Fight dataset, the Violent Crowd dataset, and the RLSV dataset, consistently surpassing existing methods in terms of accuracy. These results underscore the effectiveness of our approach in combining visual and motion-based data to improve violence detection.

In the following sections, this report will delve into the specifics of the framework's design, implementation, and evaluation. We will also discuss the limitations of the current approach, propose areas for future research, and explore the potential impact of this technology on society. Through this comprehensive analysis, we aim to contribute to the ongoing development of more effective and responsible violence detection systems, ultimately enhancing safety and security in various environment. Introduction to the Host Organization.

1.1.1 Identity of the Organization

The SETIME laboratory, based within the Faculty of Sciences at Ibn Tofail University in Kenitra, is dedicated to organizing and promoting research in the fields of fundamental and applied sciences.

1.1.2 Mission of the Organization

The mission of the SETIME laboratory is to conduct research in predefined thematic areas, in line with the four-year accreditation of Ibn Tofail University in Kenitra. The laboratory undertakes research projects, facilitates collaboration among researchers, hosts researchers and Master's students, promotes cooperation among its members, disseminates scientific output, and organizes scientific events. In addition to fundamental and applied research, it may also engage in R&D activities and provide services related to its areas of expertise.

1.1.3 Research Areas

The SETIME laboratory is comprised of four research teams working on the following topics:

- Electronic Systems and Telecommunications
- Renewable Energies and Materials Engineering
- Mechanical Engineering
- Information Processing and Artificial Intelligence

2.1 Introduction to the topic

Violence is a pervasive issue across the world and very subjective concept manifesting in various forms such as physical, psychological, and verbal aggression, as depicted in Fig.1. However, ‘‘Violence is often understood as an act of physical force or power that either results in or has a high probability of causing injury, death, psychological harm’’. The global impact of violence is substantial, with over 1.6 million deaths annually and have significant implications for social economies, including healthcare expenses, security measures, police enforcement, and property damage.

Initially, the rise of violent incidents has prompted the installation of security cameras in public places, such as schools, airports, hospitals and shopping malls to ensure the security and safety of people’s life. It can assist the authorities by alerting them about violent behavior while monitoring people’s behavior. However, the responding time of human operated monitoring system was extremely slow, resulting loss of person’s life and property. Also it has become nearly hard to detect violent behavior manually due to the tremendous rise of global population and exponential growth of surveillance equipment’s in contemporary era. Hence, the identification of violent incidents has emerged as a prominent research area in computer vision over recent years due to its application and recent technological advances. Violence detection represents a significant computer vision challenge, with the goal of automatically and efficiently determining whether or not violence occurs within a brief time frame.

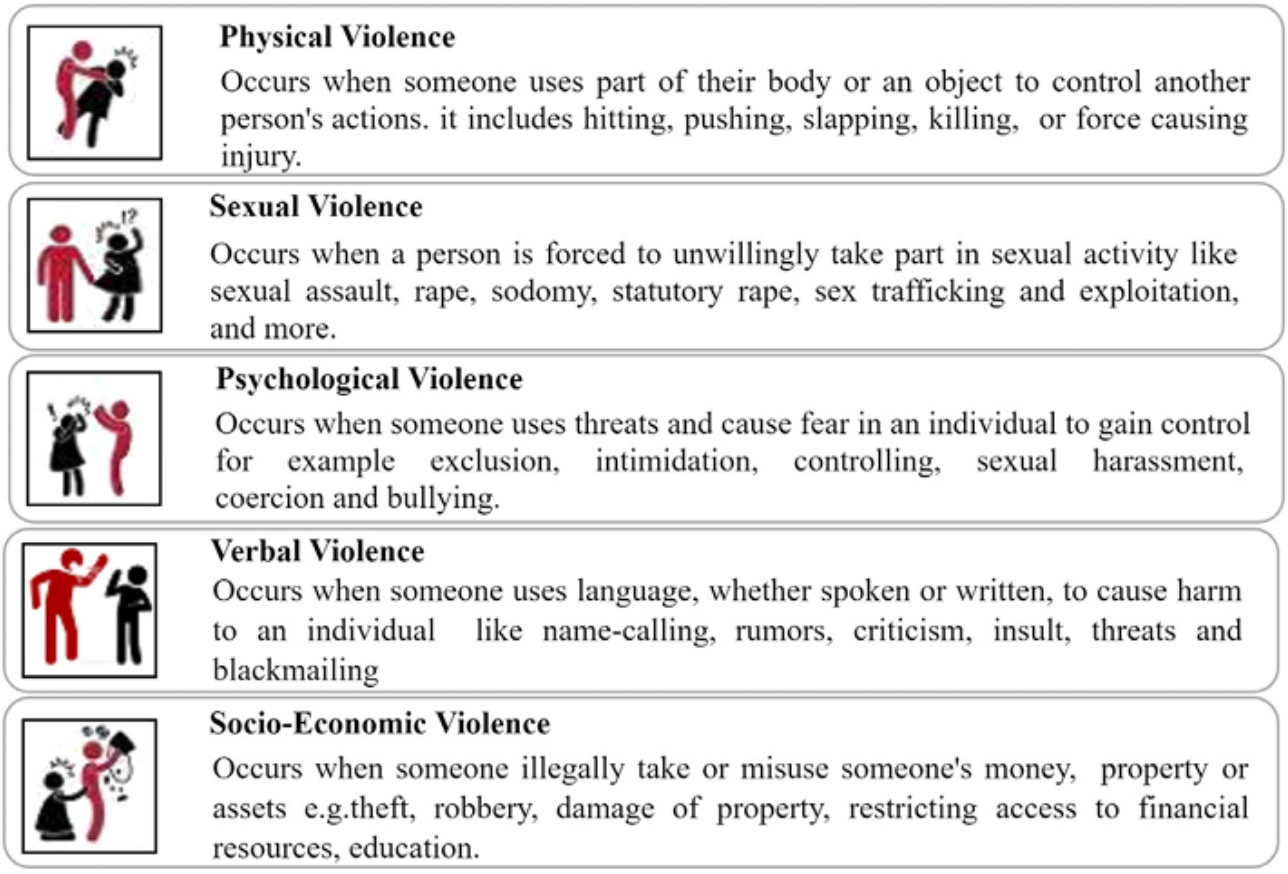


Figure 1 : TYPES OF VIOLENCE

2.2 Motivation

Violence detection in videos is motivated by a commitment to enhancing public safety and promoting societal well-being. As violence remains a persistent global issue, detecting it in real-time has far-reaching implications for both law enforcement and public security. The inherent complexity of violence detection, coupled with the intellectual pursuit of developing robust algorithms, drives researchers to explore innovative and precise solutions. Accurate violence detection not only strengthens community safety but also fosters a proactive response to criminal activity and dangerous situations.

To effectively address this global challenge, numerous researchers have proposed a wide array of advanced techniques, leveraging the power of computer vision, machine learning, and deep learning to automatically identify violence in videos without the need for human intervention. These approaches are capable of analyzing vast amounts of video data in real-time, providing valuable insights to authorities for rapid response. This capability is crucial not only in public spaces, where surveillance systems monitor potential threats, but also in areas like online video filtering, screening social media platforms for harmful or graphic content, and enforcing community guidelines.

The rapid detection and identification of violent behavior empower law enforcement and security agencies to take preventative measures, mitigating harm before it escalates. Furthermore, violence detection technologies contribute significantly to safeguarding the public through their integration into monitoring systems, ensuring a watchful eye on environments where violence may occur. By continuously refining and advancing these methods, researchers play a critical role in shaping safer communities and offering technological solutions to one of society's most pressing concerns.

Chapter 1
Literature Review

1. Introduction

The process of violence detection begins with gathering input video data containing violent and non-violent scenes from a variety of sources, such as CCTV footage, movie clips, and YouTube video. In literature, most of the studies utilized published benchmark datasets. Researchers have developed several techniques for detecting violence from videos, but a basic process of violence detection follow some common steps which includes: Pre-processing of the input video sequence for missing and noisy values, feature extraction to find behavior of person and Classification by trained classifier to label video as the violent and non-violent. Fig. 2 shows the basic model to identify violent events or aggressive behavior in video sequence.

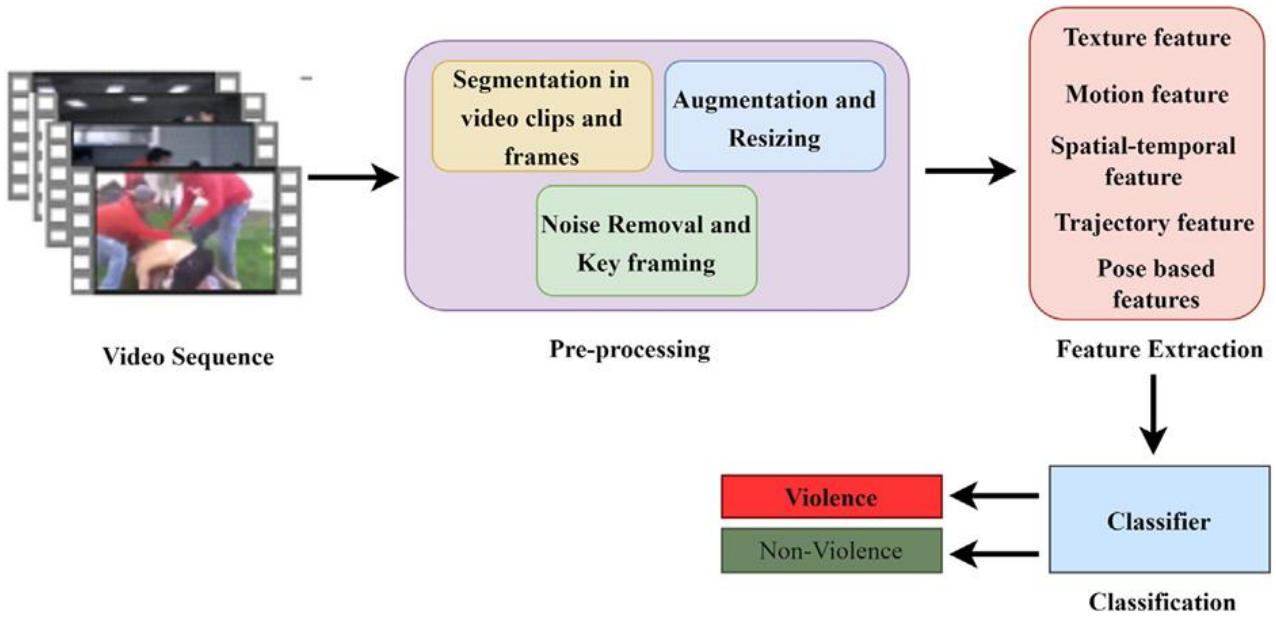


Figure 2: GENERAL FRAMEWORK FOR VIOLENCE DETECTION IN VIDEOS

The first step starts with the input of a video. To this end, a dataset of videos which contain violent and non-violent scenes is necessary. The second step is key frame extraction, although not all architectures include this step. It consists of the selection of frames that may potentially contain violence; this selection is performed in order not to have to process large amounts of video, thereby reducing the required computation levels. The third step consists in transforming the data to serve as input to the violence detection algorithm, depending on the features the algorithm is to extract. The fourth step is feature extraction and training the algorithm on these features, there are different combinations of algorithms, and therefore, there are variations in the process. Finally, a classifier decides whether the scene is violent or non-violent.

2. Basic framework for violence detection in videos

The process of violence detection begins with gathering input video data containing violent and non-violent scenes from a variety of sources, such as CCTV footage, movie clips, and YouTube video. In literature, most of the studies utilized published benchmark datasets. Researchers have developed several techniques for detecting violence from videos, but a basic process of violence detection follow some common steps which includes: (1) Pre-processing of the input video sequence for missing and noisy values (2) feature extraction to find behavior of person (3) Classification by trained classifier to label video as the violent and non-violent.

Pre-processing:

Pre-processing is the initial step during video processing. The purpose of this task is to convert raw video data into a model-acceptable format. It can be performed in two phases. First phase may involves extraction of frames from video stream, resizing the frames to fixed size and extraction of key frames by removing the redundant frames and possible data augmentation techniques like random cropping, random brightness, horizontal flipping to increase size of data . Extracted frames may contain noise caused by poor illumination and resolution. So, morphological processing may be applied to ensure the noise free frames for further processing during second phase as it enhances the reliability of blob detection of moving person in videos. There are various filters, such as Guided Image Filtering and median filtering, average filtering, bilateral filtering, and Gaussian blur to sharpen the edges of images and to improve blurry images.

Feature extraction:

Once the data is pre-processed, Feature extraction is the next step. A feature extraction is the main step for obtaining relevant information from huge data by reducing the dimensions of the data, which is used for a variety of computer vision application, such as fire detection, video summarizing, and activity analysis. In violence detection, relevant and discriminating features of image frames are identified and represented in a systematical way during this step, which can be used to describe human behavior as violent or nonviolent. These features have a significant impact on the detection of violence, so their accurate selection and representation is crucial. There are several features that are used for violence detection, such as motion, interest-point, texture, trajectory-based features, etc. In Table 1, the different kind of features and feature descriptor used for violence detection are presented.

Tableau 1: POPULAR FEATURES FOR VIOLENCE DETECTION

Feature type	Feature descriptors	Description
Motion	Optical flow, acceleration, Motion blobs, MHI	Determine the motion information of moving person's action in video
Interest Point	SIFT, MoSIFT, LaSIFT, STIP, ToRF	Detection of interest locations in both space and time domain with substantial variations in motion corresponding to actions
Trajectories	Combination of multiple descriptors(HoF, HoG, SIFT, MBH etc.)	Tracking of space interest point over time using optical-flow fields
Pose based	2D and 3D pose estimation	Tracks the spatial and temporal relations of body joint positions based on human poses
Texture	LBP, GLCM	Extract the local visual patterns such as brightness, color, shape or size

Classification:

After extracting features from videos, the next step is classification. Following the task of feature extraction, features are fed into classifiers for training, learning, and classification. Classification is the process of labeling the given set of data pattern into classes. Detecting violent and non-violent events is typically a two-class problem. In the classification step, tests videos are categorized as violent or non-violent by a classifier trained with the features acquired in the last step. The accuracy of the classification phase depends upon these features. Various Machine learning based classification models (SVM, KNN, RF, NB etc.) and deep learning based classifiers (CNN, LSTM, RNN) adopted by authors in the domain of violence detection. In previous research work, SVM is the frequently used classifier to classify violent and non-violent events.

3. Challenges in video violence detection

The objective of video violence detection is to identify, classify, and recognize aggressive behavior in video sequences, defined as violent events, to minimize or prevent hazardous situations. However, this task is challenging due to various factors that affect the detection process.

Illumination variance:

Environmental illumination varies with the brightness of the day-night light and weather conditions. A variety of illumination conditions can affect the color and contrast of outdoor video recordings, making it harder to detect violence in videos. Several researchers have tried to resolve issues related to dynamic illumination. For instance, Xu et al. [4] handled the illumination noise in their work. They used motion activation maps for active violent regions which are extracted from consecutive frames through calculating the average of all optical flow magnitudes. In order to adapt illumination changes, all areas on the map with broken pixels are removed. Proposed method by [3] utilized Kalman filter to deal with the changes in illumination and moment features to minimize the complexity of the background. In this regard, Fenil et al. [5] proposed a system based on HoG(Histogram of Oriented Gradients) function that is robust to varying illuminations and textures. In this work, images from video were segmented into non-overlapping 8*8-pixel patches called cells, and gradients were calculated for each pixel of the cells. For each cell, the gradient orientations were represented in a histogram with nine bins. Afterward, the histograms of 2*2 cells block were normalized by their neighbors to improve robustness to texture and illumination variations. In another work, authors [6] suggested to learn both appearance and dynamic information using convolution Networks to provide better tolerance for illumination variation with camera motion.

Complex dynamic Background:

Presence of changing or non-stationary background due to camera motion, movement in background such as water waves, wriggling trees, moving clouds and environmental changes often cause difficulty in target detection and feature extraction. To address this problem, Background subtraction was performed by [7]. They applied Gaussian blur on two consecutive frames to remove the background noise. Adaptive background subtraction methods based on Gaussian distributions for each pixel was adopted for dealing with lighting changes, repetitive motions from clutter, and changes in the scene over a long period of time. Another study utilized Gaussian kernel of 3 × 3 to eliminate noise and a Mixture of Gaussians (MoG) background subtraction to eliminate objects irrelevant to the actors [8]. Also, the optical flow is a convenient and widely used approach for motion estimation and representation in the presence of camera motion or changing background. Several researchers adopted optical flow based approaches [12,45,46] for complex, dynamic and noisy background. Background separation in complex scenes has been achieved by

using a fast and robust Gaussian Model of Optical Flow (GMOF) [1].

Partial/Full Occlusion:

Occlusion is perhaps the most major problem that hampers the effectiveness of behavior detection approaches in real-world circumstances. The target person can be either partially or fully occluded due the presence other objects and persons in the scene. It leads to the loss of necessary visual information such as motion and appearance. Thus, describing people and their movements for behavior analysis can be challenging because recognizable shapes change from frame to frame. The adaptability to occlusion has been reported in several studies. Llyod et al. [11] found Visual texture appropriate for representing the unstructured patterns resulting from occlusions in crowded scenes. For this task, authors extracted Haralick texture features from gray level co-occurrence matrix (GLCM) computed by counting the co-occurring gray levels intensity of a linear spatial relation between two pixels. Zhou et al. [2] suggested two features Local Histogram of Oriented Gradient (LHOG) retrieved from RGB frames and Local Histogram of Optical Flow (LHOF) obtained through motion magnitude images. In this work, initially, block strides were divided in half, resulting in overlapping half blocks. In the next step, each LHOG was normalized. Then, LHOFs were generated from the motion magnitude images to provide better adaptability to occlusion and illumination variations. To address the occlusion problem Zhang et al. [12] proposed an descriptor named Motion Weber Local Descriptor (MoWLD) which extract local appearance with an aggregated histogram of gradients from consecutive regions of the image. Due to the ability to encode both motion and appearance features of the objects in the sequential frames makes local spatio-temporal features more robust to occlusion, illumination inconsistency, and cluttered backgrounds. In this context, STACOG(Spatio-Temporal Autocorrelation of Gradient-based Features) was explored to learn local relationships among spatial and temporal gradients for a better understanding of cluttered and occluded crowd scenes [13].

Blur Motion:

Sudden or rapid motion of the object and camera could cause motion blur in image or sequence of images (video). Usually, motion blur moves the frequency spectrum of an image closer to the low frequencies. As a result, it becomes very difficult to compute optical flow and motion estimations have been considered in some works. An early study [14] proposed an efficient method for recognizing violence based on extreme acceleration patterns without using tracking or optical flow techniques. They found that extreme acceleration and global camera motion result in blurred images, making tracking more difficult or even impossible. To remove the blur caused by global motion, a deconvolution pre-processing step was performed. For each pair of frames, global motion was calculated using phase correlation. Once the global motion was computed, the length and angle of the displacement were calculated to build a Point Spread Function (PSF). This PSF was then used for the deconvolution of the next frame using the Lucy-Richardson iterative deconvolution method. Mukherjee et al. [15] found that motion patterns, such as acceleration and deceleration, were useful for tracking moving objects. First, local and global motion maps were extracted by subtracting two adjacent frames. Motion blur caused by global motion was eliminated by extracting low-frequency portions in the image, while local motion with high-intensity regions was enhanced to compute acceleration based on the displacement that occurred in each motion map. More recently, Pujol et al. [16] estimated accelerations between images based on the assumption that moving objects follow consistent patterns.

Camera Motion:

Tracking moving objects or persons with moving cameras is challenging because it involves not only the motion of the object but also the motion of the camera. As a result, estimating and compensating for camera motion is necessary to extract motion trajectories, optical flow vectors, and image derivatives from the video. Senst et al. [17, 18] observed that the motion signals of the direction field were greatly affected by camera motion. To compensate for this global camera motion, a homography-based background motion model was used, and compensation was achieved by subtracting the background from the actual direction field. In [19], an optical flow approach was employed for motion analysis, as the magnitude and direction of optical flow vectors are important measures of motion. To calculate the optical flow between two adjacent frames, a motion-resilient algorithm was applied. Camera motion usually causes the background to move consistently, while human activities tend to move in a less uniform manner. To eliminate background noise caused by camera motion, uniform motion regions were removed from the optical flow images.

4. Violence detection techniques

Detecting violent event is essential since it is directly related to safety and security in public space for the protection of the people's lives and property. Researchers and academicians have used a variety of features and approaches to detect such unusual and suspicious violent activities. These techniques has been categorized into three categories:

- The detection of violence using handcrafted feature-based representation.
- End-to-end deep learning techniques for detecting violence.
- Hybrid violence detection approaches.

4.1 Violence detection using hand-crafted features

The handcrafted feature-based representation is a traditional approach based on machine learning. To detect violence in videos, researchers have employed various types of manually designed features (interest points, motion, trajectory) to extract visual information from input frames then applied machine learning algorithms like SVM, KNN and random forest for classification and to assign labels to them. These methods can be further sub-divided based on the features they extract as reflected in Fig. 3.

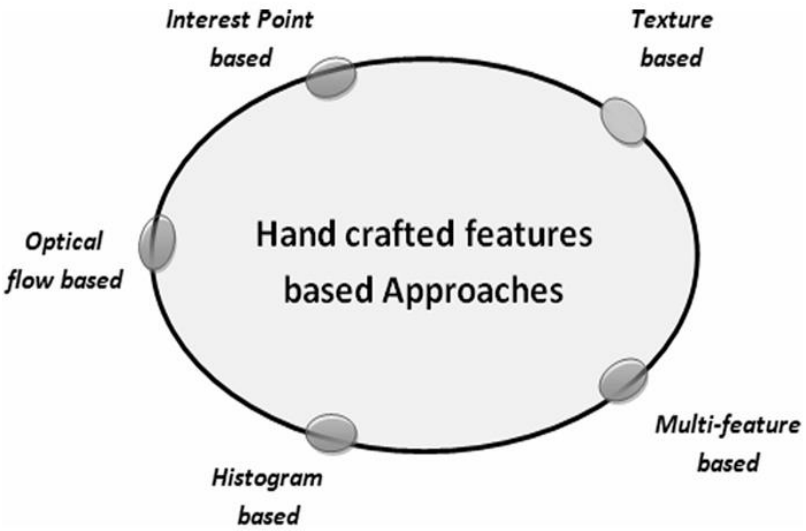


Figure 3: HAND CRAFTED FEATURES BASED VIOLENCE DETECTION APPROACHES

Since every feature extraction approach has its upsides and downsides as discussed in Table 2, it's pretty difficult to draw a fair conclusion regarding which is better among all.

Tableau 2: COMPARISON OF DIFFERENT HAND-CRAFTED BASED TECHNIQUES

Method	Upsides	Downsides
Optical flow based	• Provide global features • Measure the temporal motion of object in videos	• Expensive to compute • Large computational requirements • Sensitive to Illumination changes and background noise
Interest point based	• Provide local features • Robust to scale, rotation, and occlusion • Low computational time	• Sensitive to noise • Not suitable for multi-person/crowded scenes • Susceptible to varying viewpoint
Histogram based	• Adaptable to scale, frequency, and velocity variations • Robust to scale, rotation, and occlusion	• Not adapted to crowded scene • Sensitive to occlusion and dynamic background • Not suitable for low quality videos
Texture based	• Low computational cost • Provide real-time description • Effective at detecting changes in crowd behavior patterns	• Doesn't not consider motion information • Sensitive to noise and distortions

4.2 Violence detection using deep learning approaches

Despite their excellent performance in capturing appearance and motion features, hand-crafted approaches are still costly for real-life practical applications because they are designed for specific problems and datasets. Therefore, after 2014, researchers began adopting deep learning approaches in Human Activity Recognition. Deep Learning (DL) techniques have also impacted the field of violence detection by replacing hand-crafted feature engineering with learnable features. In contrast to traditional techniques, deep learning methods can automatically extract important video features from input video data.

Deep learning techniques pertaining to the detection of violence have been placed under three categories as discussed below:

- Convolutional neural network based approaches: A convolutional neural network is a multilayer neural network designed to recognize visual patterns directly from pixel images with little to no preprocessing. A 2D CNN can only extract spatial information, so researchers developed 3D CNNs to apply convolution to temporal sequences for video data and enabled the remarkable capability for Conv3d to learn spatial and temporal relationships.

- CNN and LSTM based approaches: In order to develop techniques for automatic analysis of violent activities, several researchers utilized the combination of CNN and LSTM. Such approaches employ the Convolutional neural network as a spatial features extractor, then the extracted features passed into LSTM Layer to learn the temporal relation. The incorporation of both features then transit to classification layer for final result.

- ConvLSTM network based approaches: A major limitation of LSTMs when handling spatiotemporal data is that full connected layers of LSTM capture the temporal changes but failed encode changes in spatial information. To overcome this problem, ConvLSTM extends fully connected LSTM by adding convolutional operators(structures) in both input-to-state and state-to state transitions. Hence, ConvLSTM is becoming increasingly popular because it is capable of capturing spatial and temporal correlations consistently.

Within computer vision, deep learning techniques are widely employed. However, selecting optimal and critical approaches plays a crucial role in detecting violence in videos. In order to do so, various researchers adopted different architectures for extracting spatio-temporal features as explained previously. Table 3 presents the advantages and disadvantages of violence detection techniques using deep learning.

Tableau 3: COMPARISON OF DIFFERENT DEEP LEARNING BASED TECHNIQUES

Method	Upsides	Downsides
Convo3d	• Fast to learn spatio-temporal feature at once • Extract high level features • capability to capture local dependencies	• High computational complexity • Excessive memory requirements
CNN-LSTM	• Applicable with small data • Low computing power(in terms of design)	• Time-consuming to extract features due to CNN and LSTM sequence processing
ConvoLSTM	• Capable to learn long temporal dependencies • Better to handle spatio-temporal co-relations	• High computational cost • Large memory consumption

4.3. Violence detection using hybrid approaches

A hybrid approach for detecting violence relies on two independent modules: feature extraction and classification. A combination of hand crafted features with deep networks or integrating learned features with traditional hand-crafted methods may yield additional and comprehensive information to enhance violence detection.

.

Tableau 4: EXISTING WORKS ON VIOLENCE DETECTION IN VIDEOS USING HYBRID FRAMEWORKS

Year	Technique	Features	Datasets	Evaluation metrics	Remarks
2016	Three streams + LSTM	Spatio-temporal, Acceleration	Hockey Fight	Accuracy Precision Recall	By exploiting the acceleration feature, generate more relevant representations to determine violence.
2017	FightNet	Motion, Appearance, Acceleration	Violent interaction	Accuracy	The FightNet achieved greater accuracy at an acceptable computational cost
2018	Hough Forests + 2D CNN	Motion, Appearance	Movies Hockey fight Behave	Accuracy	Framework is computationally very efficient, permitting the it to be used in real-time
2019	Multi-stream VGG16	Spatial, Temporal, Rhythmic and Depth information	Hockey Fight Movie	Accuracy	It is not enough effective or efficient technique because of its too many parameters and extensive training
2019	Two stream 3D CNN	Motion acceleration	Fight Action Detection	Accuracy AUC	The motion acceleration features provide a more accurate representation than traditional optical flow while detecting violent actions
2020	C3D+SVM	Acceleration information	hockey Fight Crowd violence	Accuracy AUC-ROC	This method has good generalizability and usable in different environments
2022	Multi-stream CNN	Movement Speed, Appearance	Movies Hockey fight Violent Flow	Accuracy Recall F-score	Using long-term temporal dependency, it describes the actions and provides outstanding results
2023	AttentionC3D+ES	Motion regions	UCF- Crime	Accuracy Precision Recall	This model is computationally efficient that achieved linear time and space complexity by using the attention mechanism and genetic algorithm

5. Related work

The state of the art can be categorized into several distinct groups based on the algorithms employed for violence detection. This categorization includes seven main categories that encompass various approaches. The primary categorization is based on the algorithm type, including manual feature extraction, CNN, LSTM, and transformers. Additionally, there are specific categories for skeleton-based and audio-based methods, which focus on detecting human movements in videos or leveraging alternative data dimensions, such as audio. Moreover, a separate category has been established for the numerous studies that integrate CNN and LSTM techniques due to their widespread use.

- CNN + LSTM :

Talha et al. [20] developed a very fast and efficient real-time violence detection system, which was tested on the personal devices of the system developers. It is based on a CNN which extracts spatial features. These features feed an LSTM. The CNN also uses two fully connected layers as a classifier. Madhavan [21] developed a model, without presenting its performance results. The model was intended to address the challenges associated with classification in inconsistent weather and illumination conditions. This approach

also potentially solved the problem of dedicating a low amount of pixels for video classification. Ullah et al. [30] used Mask R-CNN for feature frame selection, which is an extension of the Faster R-CNN model based on object detection (in this article it was used to detect people and cars). It has the added ability of detecting and labeling segment objects in an image. For feature extraction, the authors used two CNNs: DarkNet and a CNN which receives as residual input optical flow. With these extracted features, a multilayer long short-term memory (M-LSTM) was fed. Vijeikis et al. [31] generated a model based on a CNN and LSTM, which made it computationally light and fast; nevertheless, it had slightly lower results in terms of accuracy. Halder and Chatterjee [32] used a lightweight convolutional neural-network-based bidirectional LSTM to identify violent activities with excellent results. Traoré and Akhloufi [33] used a pre-trained VGG-16 with INRA person dataset to extract spatial features, which then feeds a type of RNN called BiGRU (bidirectional gated recurrent unit). Finally, as a classifier the authors used three fully connected layers, the last one with softmax activation. Similarly, Ref. [34] used VGG-16 for spatial feature extraction, and a LSTM was used for temporal feature extraction. Asad et al. [35] proposed a violence detection model that took as inputs two video frames at times t andt+1(not the subtraction of both). Two pre-trained CNNs were used, one receiving frame t and the other t + 1. They were used to extract high- and low-level features. Additional wide dense residual blocks (WDRBs) were used to learn these combined feature maps from the two frames, and then, the LSTM network learned the temporal patterns between the features extracted by the CNNs. To raise an alarm that violence detection is occurring, a real-time graph was plotted with the level of violence obtained from the algorithm output and, above a certain value, the scene was identified as violent. Contardo et al. [93] used a CNN pre-trained with ImageNet called MobileNetV2, which extracted spatial features and whose output was fed into two LSTMs to test which one obtained better results: a temporal Bi-LSTM and a temporal ConvLSTM.

Gupta and Ali [36] used the VGG-16 pre-trained network whose output was fed into an LSTM and a Bi-LSTM, in order to test which of the two yielded better results. Islam et al. [37] added a multitude of parameters describing the datasets used (total classes, number of videos per class, frames per second, video length, avg. frames per video, resolution, number of locations). It focused on sexual assault detection, but also on physical assault detection. The authors used pre-trained VGG-16 and VGG-19, whose output they fed into an LSTM. Jahlan and Elrefaei [38] selected frames randomly from each frame packet, furthermore they converted the images to squares by choosing any area of the data frame in between. As input of the CNN, the selected frames were not inserted, but the difference between frame i and i + 1 was inserted. The CNN that was used was automated mobile neural architecture search (MNAS) which is a lightweight CNN. A convolutional LSTM was used not only on temporal extraction; instead, the convolutional layer added a spatial element. Three different classifiers were used to see their effectiveness; prior to the classifier, the results of the trained features were scaled between 0 and 1 (normalized). The best performing classifier was the SVM. Mumtaz et al. [39] used VGG-19 (pre-trained CNN) whose output they fed into an LSTM. At the time of publication of the paper, and to the best of the authors' knowledge, they were the first to introduce a process control technique: control charts for surveillance video data analysis. The concept of control charts was merged with a novel deep-learning based violence detection framework. Sharma et al. [40] used a CNN pre-trained with ImageNet for spatial feature extraction called Xception, the output of which is fed into an LSTM. Singh et al. [41] separated the state of the art between motion-based, machine learning, and deep learning. They used two CNNs to extract low-level features and local motion features. The result was passed to an LSTM that learned global temporal features. It was mentioned how the extracted features included edges or lines of objects and people, and body motions. These features also included appearance-invariant features such as changes in illumination, weather, and other environment or background related changes. Srivastava [42] proposed two different algorithms. One of them was a violence detection algorithm. The second one was a facial identification algorithm, useful in case violence occurs. It focused on violence

detection using drone cameras. Spatial features were ex tracted with a CNN block using architectures that had been pre-trained with ImageNet and whose output was fed into an LSTM. In total, seven different algorithms were used in addition to three combinations of some of them. Traoré and Akhloufi fed two CNNs called EfficientNet, which were pre-trained on ImageNet, one of them with optical flow and one of them with RGB. The outputs of these two CNNs were fed into an LSTM, and finally, to a classifier, consisting of an FCL with a sigmoid activation layer. Islam et al. [43] proposed a two-input structure that combined a CNN with a separable convolutional LSTM (SepConvLSTM). One of the inputs receives as input the RGB video with background suppression and the other one the frame difference (difference between frame i and i + 1). They also implemented three versions of the proposed architecture in which the classification functions vary, as well as the information that is passed to the fully connected layers. Mugunga et al. [44] decided to introduce optical flow images which, although they did not contain all the scene information, helped to reduce the computational cost and to extract spatio-temporal features. This output fed a Bi-ConvLSTM that was able to extract short- and long-term features, obtaining better results than unidirectional ConvLSTMs.

- **CNN:**

Mahmoodi et al. [45] presented a work where an image segmentation method called SSMI was used to avoid introducing all the frames of the video to the CNN. Then, with a single 3D-CNN structure it performed spatio-temporal-focused feature extraction and used fully connected layers as the classifier. Ahmed et al. [46] used a CNN-v4, highlighting that, while other state-of-the-art work used CNNs for violence detection, it introduced all the frames of the video, which was very expensive computationally, and therefore, the selection of characteristic frames for much lighter computation was essential. Ji et al. [47] presented a new dataset called Human Violence Dataset containing 1930 video clips from movie promotion videos on YouTube. However, it did not consider only physical aggression, but also gun violence. The two-stream CNN model is a neural network architecture that uses two independent streams of visual information: one for spatial information and one for temporal information. The temporal stream receives a single image, while the spatial stream receives 10 frames of optical flow. After extracting features with a CNN, a machine learning algorithm is trained to quantify the violence in the videos by optimizing the weights. The article quantified levels of violence in the processed videos through feature calculation, which it based on a ranking score. Three levels of violence (L1, L2, and L3) were determined using a confusion matrix. Ehsan et al. [48] highlighted the advantages of CNNs, where feature extraction and feature training was performed within the CNN architecture itself and not in separate algorithms. The proposed CNN architecture was called Vi-Net, which performed classification in two fully connected layers, the latter of these with the softmax activation function. Jayasimhan et al. [49] proposed a 3D-CNN followed by a 2D-CNN. It did not require transfer learning (pre-trained network) or manual feature extraction, which made the structure lightweight. The 3D layer could capture temporal information over time. The 2D layer aimed to fuse temporal features into a 2D representation. Kim et al. [50] proposed several methods to improve people tracking, as current methods still present difficulties. A 3D-CNN structure was proposed for fall detection and violence detection. Monteiro and Durães [51] used the AVA dataset, selecting 17 violent actions out of 80. Their model had two paths: a slow one for appearance details and a fast one for dynamic motion. Each path used a ResNet. Features from both paths were combined using global average pooling and a fully connected layer, followed by softmax. Then, fine-tuning was performed with the X3D network to refine the results, also resulting in a fully connected layer. Talha et al. [52] divided the related work into deep learning, supervised and unsupervised learning, knowledge distillation, and multimodal learning models. The authors proposed a model using a C3D with a classifier from a fully connected layer, with the sigmoid activation function for the last layer. Appavu [53] used a keyframe selection method for each video sequence: one frame for spatial, three for temporal and six for spatio-temporal. Multiple inputs were introduced to the

CNNfor the spatial, temporal, and spatio-temporal branches: grayscale equation, optical flow constraint equation, and differential kinetic energy image (adaptive mean thresholding [AMT], adaptive Gaussian thresholding [AGT]). It is important to highlight the use of gradcam (widely used XAI technique) to illustrate the input of the spatio temporal branch of the CNN, although the algorithm only used it to facilitate the reader's understanding and not for a real purpose of explainable artificial intelligence. Adithya et al. [54] used a pre-trained CNN for feature extraction to improve the final weights. It obtained good results. Several combinations of CNN (5, 8, and 10 convolutional layers) were made, where the 8-layer combination gave good results. Finally, the best combination was the five-layer 3D-CNN network. Bi et al. [55] used a selected number of relevant frames to filter out false positives (hugs or other actions that may look like violence). For feature extraction, ResNet18 was used (pre-trained CNN), which is a type of CNN that is easier to train and has fewer parameters than CNN-3D and LSTM. To facilitate feature extraction, an image segmentation method (Deeplab-V3plus) was used. It was argued that reducing the dimensionality of the images extracted from the datasets would allow a much more clustered feature space. Image segmentation could focus on the explainability of the algorithm, although this approach was not discussed in the paper. Chen et al. [27] used the detection of changes in brightness as a method of extracting characteristic frames. This is because the difference in brightness between adjacent images could indicate a change in the action of the video. Where the maximum variation in brightness determined segmentation. As an algorithm for feature extraction and training, a combination of pre-trained networks between ResNet and Inception-V1 was used. Freire Obregón et al. [56] used YOLOv4 (which detects and labels an object in an image) and SiamRPN (which focuses on tracking an object) for the extraction of characteristic frames. For violence detection, a 3D-ConvNet with two inputs was used, which was also pre trained by the Kinetics dataset algorithm. In addition, different classifiers were evaluated to assess which one obtains the best accuracy with the proposed structure. The effect of the environment on the accuracy of the algorithm was analyzed, showing that footage without context produces a deterioration of the classifier between 2 and 5%. The authors also showed that performance stabilizes for context-free sequences, regardless of the level of context constraint applied. In addition, the article evaluated training and testing with different combinations of the datasets used. It was observed that better results were obtained when using the same dataset for training and testing. Gkountakos et al. [57] used a pre-trained 3D-ResNet with fully connected layers for classification. Huszár et al. [58] proposed two pre-trained architectures: Fine-tuned X3D-M model and transfer-learned X3D-M model. The fine-tuned X3D-M model optimized the X3D-Mparameterslearned from the Kinetics-400 dataset, while the transfer-learned X3D-M model extracted spatio-temporal features first, without modifying the X3D-M parameters, to train multiple fully connected layers. Better results were obtained with the fine-tuned X3D-Mmodel. Jain and Vishwakarma [59] used dynamic image as input to the algorithm, which focused on the motion of salient objects in the video, combining and averaging background pixels with motion patterns while retaining long-term kinetics. Features were extracted with ResnetV2 pre-trained with ImageNet. Training was performed in the final layers of this algorithm along with fine-tuning techniques (freezing certain values of the neurons to adjust the weights). Liang et al. [60] used YOLOv5 for the extraction of characteristic frames. YOLOv5 is an object detection algorithm, in this case, it was used to detect people in video. DeepSort was then used to track the person. Subsequently, a SlowFast network with ResNet pre-trained as the CNN was used for feature extraction and training. One of the challenges to be solved in violence detection involves the difficulty in identifying the different positions and angles of the subject with respect to the surrounding space. The following steps were used to solve this challenge: the exact positions of the subjects were determined in each video frame, the movement of the subjects was tracked over time, the location where violent actions occurred was identify and marked, these coordinates were mapped onto a real geographic space, and the spatial and temporal information from the video was fused, to represent violent behaviors on a geographic map, allowing for an accurate understanding of their location in the real world. Mumtaz et al. [61] proposed a violence detection method that used an

architecture called Deep Multi-Net (DMN), which combined pre-trained convolutional neural networks (CNNs) such as AlexNet and GoogleNet with ImageNet. Violence detection was performed on sequences of images or frames taken from field hockey and movie datasets. DMN not only excelled in terms of accuracy outcomes, but it also demonstrated superior efficiency with rapid learning abilities, surpassing both AlexNet and GoogleNet, with a remarkable 2.28-fold speed advantage, all while maintaining comparable competitive accuracies. Que et al. [62] focused on the detection of violence in long-duration videos using pre-trained CNNs. The authors focused on achieving accuracy in determining when the episode of violence began and ended since it is an understudied area in the literature. During the second phase, the deconvolution technique was employed to precisely pinpoint the potential video segment down to the individual frame, thereby establishing the exact moment of violence occurrence. It was emphasized that the preprocessing could be much better, as well as the method of detecting the start and end of the violence in the video. Santos et al. [63] asserted that 3D-CNNs are more advanced than conventional CNNs, being able to extract temporal as well as spatial information. The authors used a CNN pre-trained with Kinetics-400 called X3D neural network. Sernani et al. [64] repeatedly mentioned the importance of robustness on violence detection and how the AIRTLab dataset was specifically designed to test the robustness of the algorithms. The paper focused on testing whether its algorithms exhibited these characteristics or not. Three different DL algorithm structures were presented. The pre-trained 3D-CNNs were trained with the Sports-1M dataset. The authors highlighted that transfer learning can improve efficiency, rather than algorithms trained from scratch, and cited other work to prove this. Shang et al. [65] divided the related work into deep learning models, supervised and unsupervised learning, knowledge transfer (knowledge distillation), and multimodal learning. Firstly, the authors proposed to transfer information from large datasets to small violent datasets based on mutual distillation with a pre-trained self-supervised model for RGB vital features. Second, the multimodal attention fusion network (MAF-Net) was proposed to fuse the obtained RGB features with stream and audio features to recognize violent videos with multimodal information. Magdy et al. [26] separated the video into frame packets and an optical flow-based technique was applied to detect areas with motion. The paper compared CNN-3D versus CNN-4D architectures. It stated that CNN-3D was good at analyzing short time spans, but CNN-4D also allowed for understanding more complex spatio-temporal relationships. The CNN had been pre-trained using ImageNet. Huaet al. [66] proposed incorporating a residual attention module into the stacked hourglass network to improve human pose estimation by addressing the reduction in initial image resolution. The new architecture enhanced the resolution and accuracy of image features. The experimental results demonstrated that this approach achieved more accurate and robust human pose estimation in images with complex backgrounds. Liu et al. [67] proposed a novel method for maintaining temporal consistency in human pose estimation from video using structured-space learning and halfway temporal evaluation methods. By employing a three-stage multi-feature deep convolution network, the method ensured accurate and stable human pose estimation with superior long-term consistency across video frames.

- Manual Feature-Based Approach :

Wintarti et al. [25] selected 20 frames from each video sequence randomly as feature frame selection. For feature extraction, two methods were used: principal component analysis (PCA) (dimensionality reduction) and discrete wavelet transform (DWT), which consists of a set of sub-bands of signal frequencies for video processing. A support vector machine (SVM) algorithm was trained from the extracted features and the SVM itself acted as a classifier. Mohtavipour et al. [68] divided the related work into deep learning, global handcrafted, and local handcrafted. The energy difference was used as input to the algorithm. A CNN was trained that had three inputs. For each sequence, one frame was selected for the spatial, three for the temporal, and six for the spatio-temporal. The spatial focused on appearances and was converted to gray video, the temporal used optical flow, and the spatiotemporal generated a differential motion energy image.

Lohithashva [29] used local orientation pattern (LOOP) as a manual feature extraction technique. These features were trained by an SVM which also acted as a classifier.

Jaiswal [69] used local binary pattern (LBP) and fuzzy histogram of optical flow orientations as manual feature extraction techniques. As a feature training technique it used AdaBoost (adaptive boosting) which is a machine learning technique. Finally, as a classifier it used a technique based on decision trees called Ensemble RobustBoost aggregation. Huetal. [70] proposed two architectures. Starting from a video clip, the authors divided the video into three 3-dimensional arrays: X-T plane, Y-T plane, and X-Y plane. TOP ALCM was then applied, which is an image processing and data analysis technique used to capture co-occurrence patterns of features in different directions and angles in a matrix. TOP-ALCMs are a matrix representation of these co-occurrences. From the TOP-ALCM results, two architectures were proposed, either the use of an SVM from the obtained elements such as entropy, energy, etc.; or the use of a CNN for training and classification.

- Skeleton-Based Approach :

Zhou [78] divided related work between human pose estimation (skeleton keypoints) and action recognition (where DL and traditional methods were introduced). In the initial feature extraction, each video frame was passed through a 3D convolutional neural network (3D-CNN), considering both spatial and temporal information across frames to capture action features. Then, features extracted by the 3D-CNN, such as HRNetW32, were divided into smaller patches to process manageable portions and capture fine details. Afterwards, each extracted feature patch was flattened into a one-dimensional vector, and position embeddings were added to account for positional relationships. Then, the TokenPose Model was used, where the sequence of patch vectors, along with special tokens representing keypoints in video actions, was fed into a transformer layer, capturing relationships between patches and keypoints. Finally, a multilayer perceptron (MLP) in the model's head predicted keypoint heatmaps, using tokens from the final transformer layer as input, ultimately locating relevant keypoints in the video, followed by classification based on spatio-temporal features to determine if the video contained violent content. It is worth mentioning that saliency maps were applied to show how the algorithm selected the key points (articulations) of the people appearing in the video. This served as an explanation for the reader.

Hunget al. [79] divided the related work into deep learning models, supervised and unsupervised learning, knowledge distillation, and multimodal learning. They used a deep-learning-based method to obtain, from the introduced video, the number of skeletons, the distance between skeletons, and the changes in human skeleton hand acceleration. From these extracted features, an SVM was trained for the classification of violent and non-violent scenes. Naik and Gopalakrishna [71] made a different division of the related work: optical-flow-based methods, histogram of optical flow, space–time interest-point based methods, and convolutional neural network methods. DeepPose was used for the extraction and training of the body positions of people appearing in the video, whose output fed an LSTM that extracted temporal relations.

Narynov et al. [72] classified the types of skeleton-based algorithms on RGB images into two categories: top-down algorithms and bottom-up algorithms. They used a pre trained CNN called PoseNet to extract and train the body position of people appearing in the video. From the detected objects (people), skeleton features were extracted, and a temporal tracking of the extracted skeletons was carried out. The detection involved a distinction between punching and kicking. Srivastava et al. [73] proposed violence detection in a new dataframe created with images taken by a drone at a certain height from the ground. The paper involved identifying the human figure, selecting keypoints in their posture by using a CNN with two inputs. From those extracted and trained features, an SVM classifier was used which divided the violence detection

into six features. Su et al. [74] extracted skeleton-based features, representing a geometric X, Y, Z map, where Zrepresented the temporal dimension. It was able to differentiate “heads” so that it could track the movements being made. These geometrical maps positions were taken to train SPIL (skeleton point interaction learning) for classification.

- **Audio-Based Approach :**

Mahalle and Rojatkar [80] established that the main objective for using audio feature extraction was to reduce the data dimensionality by extracting the most important features from audio samples. When the feature vector dimensionality is small, a set of features can be useful for representing the characteristics of audio samples. From the audio feature extraction, labels were assigned to each instance. All the labels with their audio were introduced to an extreme learning machine which was trained to identify violent audios. Wuet al. [75] presented a large-scale multi-scene dataset called XD-Violence. The model for violence detection involved a neural network comprising three concurrent branches aimed at discerning various connections among video fragments and merging attributes. The comprehensive branch grasped extensive dependencies through similarity precedence, while the localized branch seized local positional correlations using proximity precedence, and the score branch dynamically gauged the proximity of the anticipated score.

Zhengetal. [81] for violence detection combined video analysis from non-equidistantly selected frames and audio from those videos. Feature extraction was performed on RGB video and audio using a transformer encoder architecture network (to extract deep temporal-spatial features) and VGG (to extract audio features). These extracted features were then run through an LSTM to extract temporal features. Cheng et al. [76] proposed a model called VARN, which consisted of pseudo-3D convolutional networks pre-trained with UCF101 and AudioNet (pre-trained CNN starting from SoundNet). The fused image and audio elements were passed to two fully connected neural networks. The reasoning network generated a vector of size 7X1, representing seven types of conflict events (no conflict, personal verbal conflict, personal physical conflict, personal physical conflict with weapons, group verbal conflict, group physical conflict, and group physical conflict with weapons). The so-called predicted network was responsible for predicting the degree of danger of conflict events.

- Transformer-Based Approach:

Akti et al. [82] used an algorithm called ViT. The vision transformer (ViT) is a neural network architecture that combines elements of transformer-based vision models with self-attention. This algorithm first patches the image and extracts features from each of the patches, taking into account the position of each patch in the original image. Then, that information is passed to another layer where temporal relationships between patches are extracted, and finally, a classification is performed. In addition, in this article, a dataset with images and videos obtained from the Internet was presented.

Ehsan et al. [83] approached the extraction of feature frames by detecting the people present in the video and removing all background information using YOLO. The Farneback methodwasusedforimageopticalflowcalculation. STAT,agenerativeadversarialnetwork (GAN)-based algorithm that performs the unsupervised translation of temporal motion features in video sequences to spatial image frames, was used for feature extraction. In the proposed algorithm, the STAT network consisted of a generator (G) and a discriminator (D). The generator took motion variation features extracted from video sequences and translated them into image frames. The discriminator evaluated the authenticity of the generated images compared to the real images. During training, the generator attempted to generate

realistic images, while the discriminator tried to distinguish between real and generated images. After training, the STAT generator (with the generator only) was used to translate normal motion features into normal images, but failed to reconstruct violent actions. Interpretation of the difference between the original and reconstructed images allowed for the classification of human behavior as normal or violent.

Kumaret al. [77] introduced a streamlined transformer model, drawing inspiration from the recent achievements of video vision transformers in action recognition. Spatial characteristics were captured from the input frames, and temporal relationships among the selected frames were improved through tube let embedding. These enhanced frames were then processed using various transformer layers. Typically, despite the known requirement of extensive training with large datasets for transformer models, it has been demonstrated that effective training on relatively compact datasets can be achieved by employing efficient preprocessing techniques. In addition, the authors created a new violence detection data frame.

- LSTM-Based Approach:

Ullah et al. [84] focused on the detection of violence in industrial environments. An algorithm was proposed which used object detection for fragmentation. In addition, two types of neural networks, an LSTM and a GRU, were used to solve the gradient evanescence; both are recurrent neural networks (RNNs) that avoid the vanishing gradient in different ways. The authors also presented the architecture and the operation of the cloud distribution of their violence detection system.

6. Conclusion

The review of existing methods in violence detection reveals a diverse range of approaches, each contributing uniquely to the field. Appearance-based methods have demonstrated considerable advancements in leveraging visual features for detecting violent actions, utilizing both RGB and optical flow information. While these methods excel in recognizing detailed motion features, they often struggle with complex scenes and varying environmental conditions.

Recent innovations in hybrid methods, which integrate skeleton-based and appearance-based techniques, show promise in addressing some of these limitations. Skeleton-based approaches offer an effective means of analyzing human poses and movements, providing valuable insights into the dynamics of violent behavior. By combining these with appearance-based methods, hybrid approaches enhance detection accuracy and robustness, benefiting from the complementary strengths of both techniques.

In addition, efforts to improve computational efficiency and reduce model complexity are noteworthy. Techniques such as lightweight CNNs and DenseNet architectures have made strides in minimizing processing time and resource usage, making violence detection systems more practical for real-world applications.

Overall, the integration of skeleton-based and appearance-based methods represents a significant advancement in the field, offering a more comprehensive and effective approach to violence detection. The insights gained from this review provide a solid foundation for understanding the proposed method, which aims to build upon these advancements to further enhance detection performance and adaptability.

Chapter 2
Methodology

1. Introduction

In order to meet the real-time requirements of the algorithm, we have designed a flexible real-time violence detection framework, as shown in Figure 4. The basic idea of the method is to determine keyframe information by introducing human skeleton feature information, using YOLO-Pose as a pose extractor to construct a human action skeleton map, and combining it with real-time 2DCNN network to classify action videos. Following feature extraction, we employed an attention mechanism to emphasize the key information, and then utilized an LSTM for classification; in the following text, we will provide a detailed description of the network architecture along the process and each component.

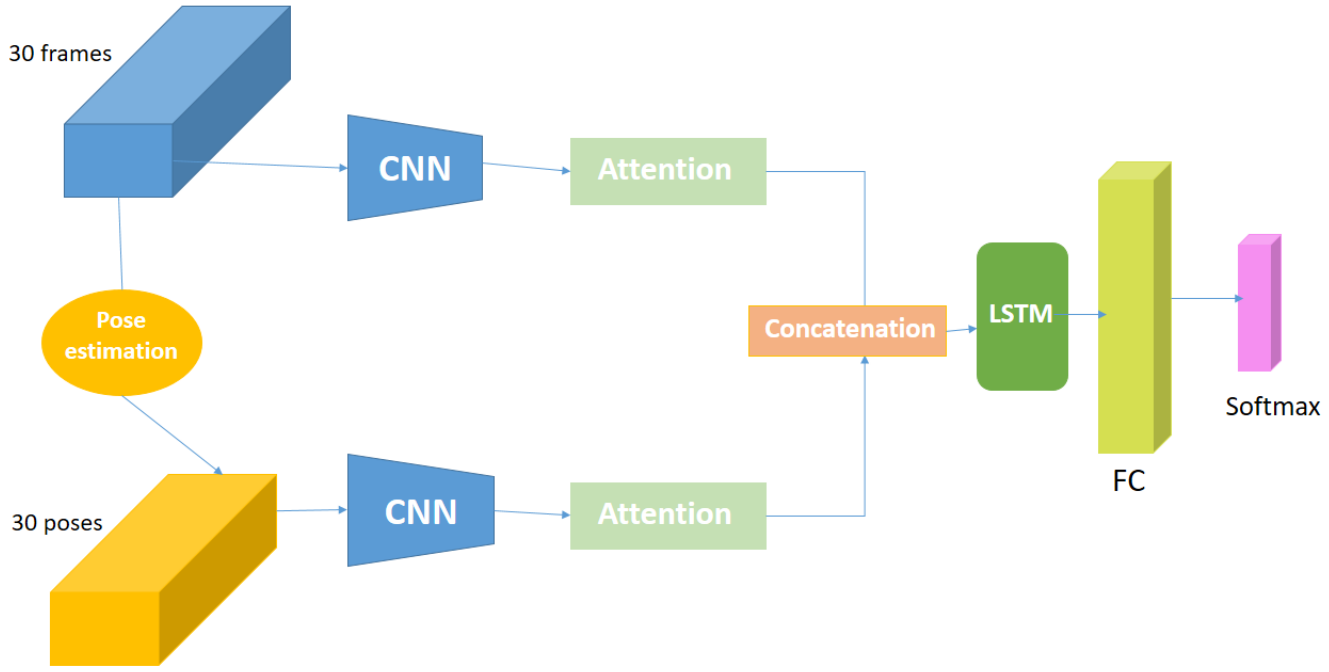
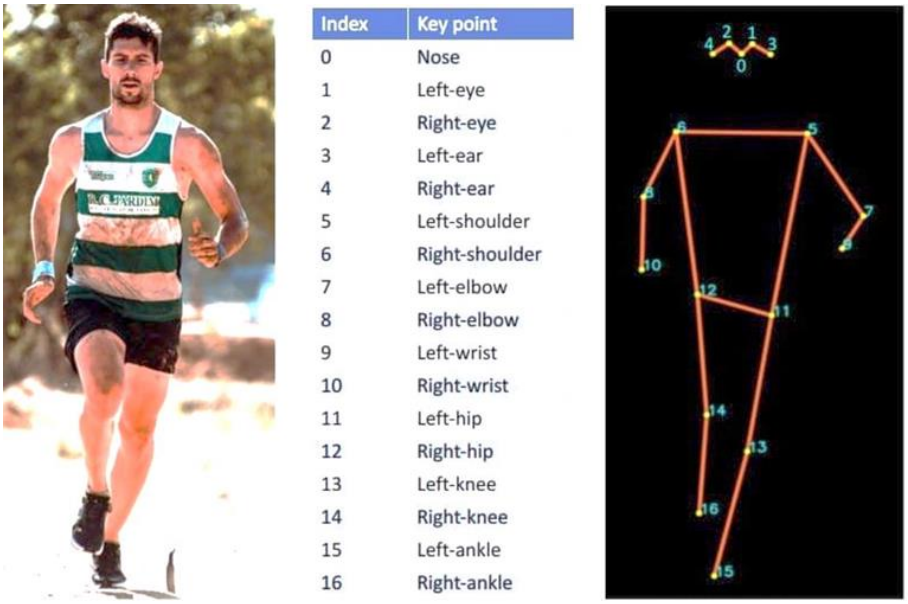


Figure 4: THE PROPOSED ARCHITECTURE

2. Human Skeleton Map Extraction

Pose estimation is a task that involves identifying the location of specific points in an image, usually referred to as keypoints. The keypoints can represent various parts of the object such as joints, landmarks, or other distinctive features. The locations of the keypoints are usually represented as a set of 2D [x, y] or 3D [x, y, visible] coordinates. The output of a pose estimation model is a set of points that represent the keypoints on an object in the image, usually along with the confidence scores for each point. Pose estimation is a good choice when you need to identify specific parts of an object in a scene, and their location in relation to each other.

Figure 5: RESULTANT IMAGE FROM YOLOV8 POSE ESTIMATION



Index	Key point
0	Nose
1	Left-eye
2	Right-eye
3	Left-ear
4	Right-ear
5	Left-shoulder
6	Right-shoulder
7	Left-elbow
8	Right-elbow
9	Left-wrist
10	Right-wrist
11	Left-hip
12	Right-hip
13	Left-knee
14	Right-knee
15	Left-ankle
16	Right-ankle

Figure 6: YOLOV8 POSE ESTIMATION KEY_POINT

Pose estimation is a critical task in computer vision, involving the identification and localization of keypoints on objects, particularly on human bodies, within an image or video frame. These keypoints usually correspond to specific parts of the object, such as joints (e.g., elbows, knees, shoulders), facial landmarks, or other distinctive features that define the object's structure. The precise identification of these points enables a model to understand the orientation, position, and movement of an object within the scene. The keypoints are typically represented as 2D coordinates [x, y] in images or 3D coordinates [x, y, z] when depth information is available. Additionally, many pose estimation models also provide a confidence score associated with each keypoint, reflecting the model's certainty about the accuracy of the detected location.

In real-world applications, pose estimation becomes especially powerful for tasks that require tracking and understanding human movements, such as action recognition, sports analysis, human-computer interaction, and safety monitoring. For instance, in the context of violence detection, pose estimation can be a valuable tool for analyzing body postures and movements that are indicative of aggressive behavior, enabling the system to identify violent actions in real-time. By detecting the relative position of limbs or identifying sudden, unnatural body movements, pose estimation models can contribute to more accurate and nuanced detection of violence in surveillance videos or live streams.

In my research, I have utilized the YOLOv8 pose estimation model, a state-of-the-art framework known for its speed and accuracy. YOLOv8 (You Only Look Once version 8) is an advanced version of the YOLO series of models, which are renowned for real-time object detection. YOLOv8 extends this capability to pose estimation by integrating both object detection and keypoint detection into a single unified framework. This allows the model to simultaneously detect multiple objects and identify keypoints on each detected object, making it highly efficient for tasks like multi-person pose estimation in crowded environments. The lightweight architecture of YOLOv8 makes it particularly well-suited for real-time applications, such as monitoring video streams from surveillance cameras, where both speed and accuracy are paramount.

The use of YOLOv8 for pose estimation offers several advantages. Firstly, its real-time processing capability allows for the rapid detection of keypoints, making it ideal for scenarios where timely action is required, such as detecting violence or abnormal behavior. Secondly, its high accuracy ensures that even in challenging conditions, such as partial occlusion or varying lighting, the model can still accurately detect keypoints. Furthermore, YOLOv8's ability to scale across different hardware setups—ranging from high-end GPUs to resource-constrained edge devices—makes it versatile for deployment in diverse environments, from large-scale surveillance systems to mobile applications.

Pose estimation is particularly useful when precise information about the spatial relationship between keypoints is needed. For instance, in violence detection scenarios, identifying the exact positions and angles of a person's arms, legs, or head can reveal a wealth of information about their actions. An aggressive posture, such as a raised arm or an unnatural bending of joints, can be indicative of impending or ongoing violence. By capturing this detailed information, pose estimation models like YOLOv8 provide a finer level of analysis compared to traditional object detection methods, which might only classify a person as "violent" or "non-violent" without understanding the nuances of their body movements.

In addition to its application in violence detection, pose estimation has a wide range of other uses. In sports analytics, for example, pose estimation can be used to track athletes' movements and analyze their form, helping coaches identify areas for improvement. In healthcare, pose estimation can assist in physical therapy by monitoring patients' rehabilitation exercises, ensuring they perform movements correctly to avoid injury. It can also play a crucial role in enhancing human-computer interaction, enabling gesture recognition systems that allow users to control devices using hand or body movements, offering a more intuitive way to interact with technology.

One of the ongoing challenges in pose estimation, however, is achieving robust performance in highly complex scenes. Factors such as occlusion, where part of the body is blocked from view, or variability in camera angles, can make it difficult for the model to accurately detect keypoints. To address these challenges, future research may explore the use of more advanced architectures, such as transformers or graph convolutional networks, which have shown promise in improving the robustness of pose estimation models. Moreover, integrating additional data sources, such as depth information or temporal data from consecutive frames in a video, could further enhance the model's ability to track and understand movements in more dynamic and cluttered environments.

Furthermore, as models like YOLOv8 continue to evolve, the integration of pose estimation with other modalities, such as audio signals or contextual information from the environment, could open new avenues for improving detection systems. For example, combining pose data with sound cues (e.g., raised voices or sudden loud noises) could help systems identify violent situations with greater accuracy, particularly in cases where visual data alone might be ambiguous or incomplete.

In conclusion, pose estimation, particularly with models like YOLOv8, is a powerful tool for understanding human movements and actions in a wide range of applications, from violence detection to sports and healthcare. Its ability to capture detailed information about the spatial relationships between keypoints makes it invaluable for tasks that require precise analysis of body movements. As research in this field progresses, further advancements in model architectures, data integration techniques, and real-time processing capabilities will continue to enhance the potential of pose estimation in solving real-world challenges.

3. CNN Network

In the context of video analysis, utilizing NASNetMobile as part of a 2D Convolutional Neural Network (CNN) architecture offers substantial advantages for feature extraction. NASNetMobile, a lightweight yet powerful architecture, is specifically designed to achieve high performance while minimizing computational resource requirements. This efficiency arises from the application of neural architecture search (NAS), a technique that optimizes the model's structure, resulting in a network that effectively balances accuracy with resource consumption.

For video data, NASNetMobile's ability to extract rich, hierarchical features from individual frames is particularly beneficial. This capability enables the model to capture complex patterns and intricate details from the video content, which are essential for tasks such as action recognition or abnormal behavior detection. By employing NASNetMobile, we can deploy a high-performing model while maintaining a manageable footprint in terms of memory and computation. This aspect is crucial for processing large volumes of video data efficiently, especially when considering real-time applications or scenarios with limited resources.

Leveraging NASNetMobile enhances both the accuracy and robustness of feature extraction, allowing the model to identify key visual elements while adhering to practical constraints on resource usage. This becomes especially important in applications where the model must operate in real-world settings, where computational power may be limited or variable. The lightweight nature of NASNetMobile allows for smoother integration into devices with restricted capabilities, such as mobile platforms or edge devices, without sacrificing performance.

Architecture of NASNet

Neural Architecture Search Network (NASNet) was developed by the Google Brain team and features two main components: the normal cell and the reduction cell, as illustrated in Figure 7. The architecture's innovative design involves initially applying its operations on a small dataset, which allows for a more streamlined and effective transfer of its learned blocks to larger datasets, achieving higher mean Average Precision (mAP) scores. This methodology enhances the model's adaptability and performance across varied tasks and datasets.

A modified version of the drop path technique, known as Scheduled Drop Path, is implemented within NASNet to provide effective regularization, improving the overall performance of the network. In the original NASNet architecture, the number of cells is not pre-defined, allowing for flexibility in model design. Normal cells are responsible for defining the feature map size, while reduction cells reduce the feature map's dimensions in terms of height and width by a factor of two. This dual structure allows NASNet to maintain a balance between capturing essential details and reducing computational load.

Furthermore, the control architecture in NASNet, which is based on Recurrent Neural Networks (RNNs), predicts the overall structure of the network based on two initial hidden states. This controller architecture employs an RNN-based Long Short-Term Memory (LSTM) model, utilizing Softmax predictions for convolutional cell predictions and constructing the network motifs recursively. This recursive design allows for the efficient generation of complex architectures tailored to specific tasks and datasets.

In our model, NASNetMobile utilizes an input image size of 224×224 pixels, while the larger NASNet model operates with an input size of 331×331 pixels. The use of pre-trained weights from the ImageNet

dataset during the transfer learning process allows NASNetMobile to effectively detect various features, such as face masks in different contexts. This transfer learning capability not only speeds up the training process but also enhances the model's ability to generalize across diverse tasks, making it a versatile choice for various video analysis applications.

By integrating NASNetMobile into our architecture, we harness the benefits of advanced feature extraction techniques while ensuring that our model remains computationally efficient. This combination of performance and efficiency is essential for the effective analysis of video data, particularly in real-world scenarios where the ability to process information quickly and accurately can make a significant impact on public safety and operational effectiveness. As we continue to explore the potential of NASNetMobile and other advanced architectures, we anticipate further improvements in our ability to analyze complex video data, ultimately leading to more robust solutions for detecting abnormal behaviors and enhancing security measures.

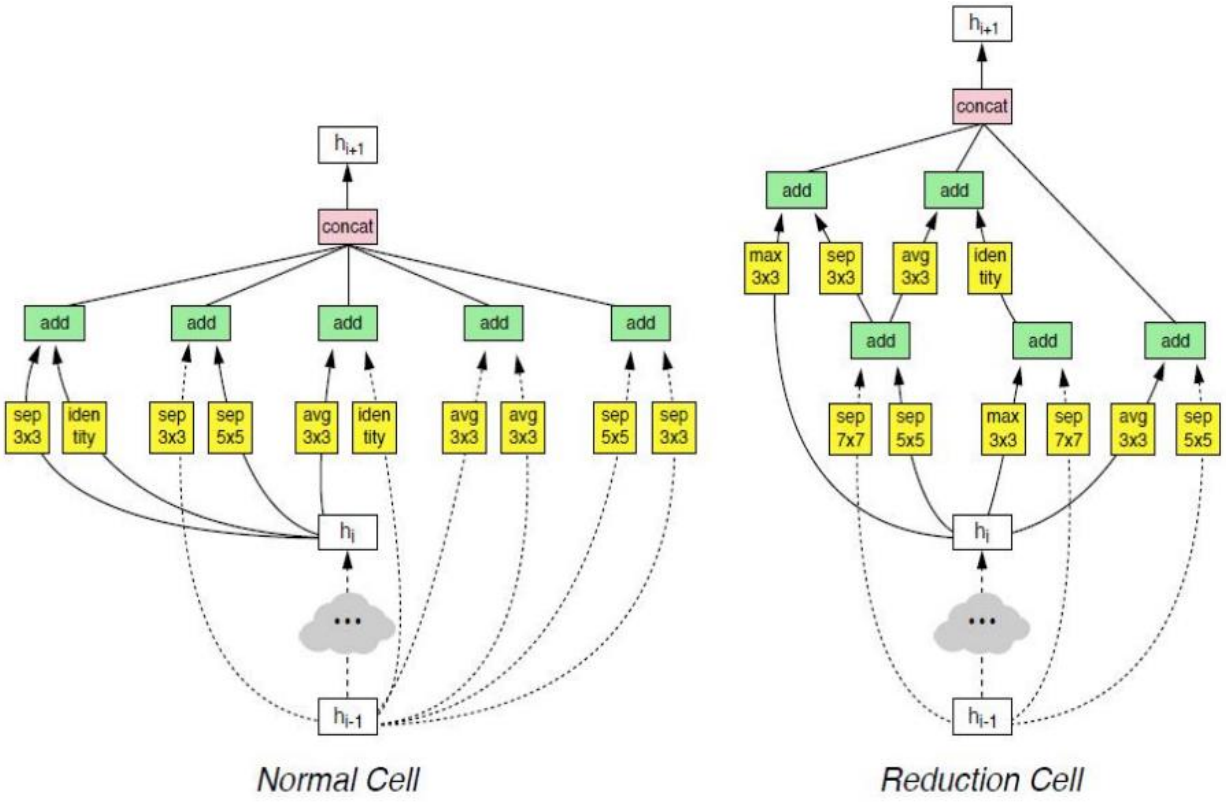


Figure 7: NASNET NORMAL AND REDUCTION CELL ARCHITECTURE

4. Attention Mechanism

After the initial feature extraction by the CNN, incorporating an attention mechanism further enhances the model's performance by allowing it to focus on the most relevant parts of the input data. In the context of video analysis, where temporal and spatial information is crucial, the attention mechanism plays a pivotal role in improving the model's ability to discern important features from less significant ones. By applying attention mechanisms, the model can dynamically weigh the importance of different regions or frames within the video, directing its focus towards the areas most relevant to the task at hand, such as detecting anomalies or recognizing specific actions. This selective emphasis helps in refining feature representations, leading to better performance in subsequent stages of the network. Moreover, attention mechanisms can mitigate the impact of noisy or irrelevant information, providing a more robust and discriminative feature set. Integrating attention layers with NASNetMobile-derived features leverages the strengths of both

architectures, resulting in a more effective and efficient model for complex video analysis tasks.

The general attention mechanism by Bahdanau et al (2014) is a powerful tool in deep learning models, particularly for tasks like machine translation, where capturing relationships between different elements in a sequence is crucial. The attention mechanism allows the model to focus on different parts of the input sequence when producing an output, helping it to consider context more effectively.

4.1. Components of the Attention Mechanism

The attention mechanism revolves around three main components:

1. **Queries (Q)**
2. **Keys (K)**
3. **Values (V)**

These components can be compared to parts of the attention mechanism proposed by Bahdanau et al., where:

- **Query (Q)** is analogous to the previous decoder output.
- **Keys (K)** and **Values (V)** are analogous to the encoded inputs.

In the general attention mechanism, the keys and values can be different vectors, unlike in the Bahdanau attention, where they are the same.

4.2. Mathematical Formulation

1. **Score Calculation:** For a given query vector qi , the mechanism computes a score with each key vector kj using the dot product:

$$\text{score}(q_i, k_j) = q_i \cdot k_j$$

2. **Softmax Operation:** The scores are then passed through a softmax function to compute the attention weights:

$$\alpha_{ij} = \frac{\exp(\text{score}(q_i, k_j))}{\sum_{j'} \exp(\text{score}(q_i, k_{j'}))}$$

where αij represents the attention weight between the i-th query and the j-th key.

3. **Weighted Sum of Values:** The output of the attention mechanism is a weighted sum of the value vectors vj, where the weights are the attention weights computed in the previous step:

$$\text{Attention}(q_i, K, V) = \sum_j \alpha_{ij} v_j$$

This equation essentially means that the final output for each query qi is a combination of the value vectors, with the most relevant ones (as indicated by the attention weights) contributing more to the output.

By using this mathematical framework, the attention mechanism effectively allows the model to focus its computational resources on the most pertinent information. This is especially critical in video analysis, where the presence of numerous frames can lead to overwhelming amounts of data. Attention mechanisms streamline this process, enabling the model to prioritize information that is likely to contribute to accurate predictions and decision-making.

In summary, the integration of attention mechanisms into the model architecture not only enhances feature extraction and representation but also fosters a more sophisticated understanding of the relationships between different temporal and spatial elements in video data. As we continue to refine our approaches to violence detection and related tasks, the attention mechanism will remain an integral component, driving improvements in accuracy and efficiency. By focusing on the right parts of the input data, we can enable models to perform with greater precision, ultimately leading to more effective solutions for complex challenges in video analysis.

5. Feature prediction network based on LSTM network

After using the temporal attention block, we get time sequence Tt . The sequence Tt implies the contribution of the features of historical moments in the same spatial position to the features of current time. The features which make more contributions to the current features will be given higher attention, and the features which make less contributions to current features will be given less attention. In order to capture the relationship between the features of historical moments and current moments, we introduce a long short-term memory (LSTM) network for feature prediction. This prediction network takes the historical features with temporal attention as input and uses the feature of the current moment as feature prediction output. LSTM is a variant of RNN, which solves the problem of gradient disappearance and inability to lose initial dependencies when dealing with long sequence. The naive RNN always accepts time step input, but the LSTM unit controls the state of its gates (input gate, forget gate, output gate) by applying an S-shaped function to them, thereby reject or accept the input. The network structure of a LSTM unit is shown in Fig. 8, and the output of each gate is shown in Equation (2)–(7).

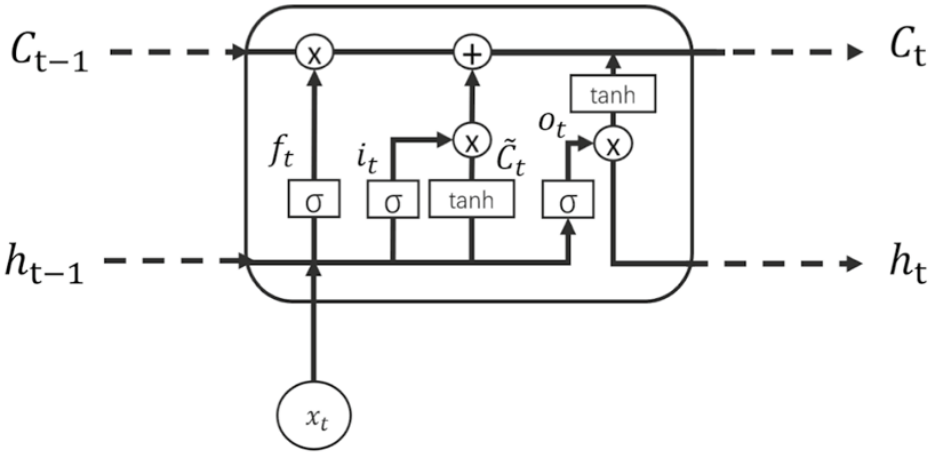


Figure 8: THE BASIC STRUCTURE OF THE LSTM UNIT

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = \sigma(f_t \times C_{t-1} + i_t \times \tilde{C}_t) \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \times \tanh(C_t) \quad (7)$$

Where W∗ represents as the weight vector corresponding to the gate, b∗ represents as the bias corresponding to the gate. We use the LSTM cell's hidden state ht as the final output. In the training phase, we use the normal behavior feature of current moment as the expected output, so that the LSTM decoding network can learn the regularity of normal behavior. In the test phase, we use the entire trained model to predict the sample's features for every moment. Those features which are significantly different from the actual features are regarded as abnormal features.

During the training phase, we utilize the features representing normal behavior at the current time step as the expected output. This allows the LSTM decoding network to learn the regularities and patterns characteristic of normal behavior. By establishing a robust model of what constitutes normalcy, we can better identify deviations from this norm. In the testing phase, the entire trained LSTM model is employed to predict the feature set for each moment in the sequence. Any features that significantly diverge from the expected outputs indicative of abnormal behavior are flagged as anomalies.

Furthermore, the integration of LSTM into our framework enhances the model's ability to understand complex temporal relationships in the data. LSTMs excel in situations where the timing and sequence of events play a critical role, allowing for improved predictions based on past inputs. This characteristic is particularly valuable in applications involving video analysis, where understanding the flow of actions and interactions over time is essential for accurate violence detection. By leveraging the predictive power of LSTMs, our model not only identifies instances of abnormal behavior but also provides contextual insights into the sequence of events leading up to those anomalies.

Moreover, LSTMs are designed to learn and remember long-term dependencies, making them well-suited for scenarios where past behavior informs current actions. This capability is critical in understanding nuanced behaviors that might evolve over several frames or time steps. For instance, in the context of violence detection, recognizing that a sudden aggressive gesture is preceded by a specific sequence of movements can greatly enhance the model's predictive accuracy. By incorporating the historical context into the analysis, the LSTM helps the model discern patterns that might not be immediately obvious from isolated observations.

In addition to their ability to capture long-range dependencies, LSTMs can be further enhanced through advanced techniques such as bidirectional processing, where two LSTMs are employed—one processing

the sequence in the forward direction and the other in reverse. This bidirectional approach allows the model to utilize context from both past and future frames, significantly improving its understanding of the temporal dynamics involved in behavioral patterns.

In conclusion, the integration of LSTM networks into our framework provides a robust mechanism for feature prediction in the context of violence detection. By effectively leveraging the temporal attention mechanism alongside LSTMs, we enhance the model's ability to understand complex relationships in sequential data, ultimately improving the identification of abnormal behaviors. As we continue to refine and expand this framework, LSTMs will play a pivotal role in our efforts to develop more accurate and reliable systems for real-time monitoring and analysis of violent incidents in video data.

Chapter 3
Experiments

1. Datasets

In this study, we employed a combination of diverse datasets specifically curated for violence detection tasks. These datasets include a range of scenarios, from controlled environments to real-life situations, providing a comprehensive basis for developing and evaluating our model's ability to detect violent behavior across various contexts.

1.1. Hockey-Fight Dataset

This dataset was proposed by Nievas et al. in 2011 and was collected from videos of National Hockey League (NHL) hockey matches. The dataset contains a total of 1000 sequence fragments (clips), which are evenly divided into two types, with 500 being 'fighting' and the other 500 being 'non fighting'. Each sequence segment lasts for approximately 2 seconds and consists of approximately 41 frames with a resolution of 360 * 288.

Figure 9: SAMPLES OF THE HOCKEY FIGHT DATASET'S FRAMES

1.2. Violent-Flows\Crowd Violence Dataset

A database of real-world, video footage of crowd violence, along with standard benchmark protocols designed to test both violent/non-violent classification and violence outbreak detections. The data set contains 246 videos with a resolution of 320 × 240. All the videos were downloaded from YouTube. The shortest clip duration is 1.04 seconds, the longest clip is 6.52 seconds, and the average length of a video clip is 3.60 seconds.

Figure 10: SAMPLES OF CROWD VIOLENCE DATASET'S FRAMES

1.3. Real Life Violence Situations Dataset

This dataset comprises 2,000 videos, evenly split between 1,000 violent and 1,000 non-violent clips, all sourced from YouTube and real-life human activities. The videos are presented in a resolution of 480 × 720 pixels, capturing scenes in both indoor and outdoor environments. The violent videos depict real street fights under various conditions, while the non-violent videos feature a wide range of everyday human activities, including sports, eating, and walking. The primary objective of this dataset is to support research and development in violence detection.

Figure 11: SAMPLES FROM REAL-LIFE VIOLENCE SITUATIONS (SOLIMAN ET AL. 2019)

2. Evaluating Indicator

We followed the original method and selected Precision, Recall, and Accuracy as evaluation indicators. Accuracy is usually used to evaluate the accuracy of detection tasks, and the larger its value, the better the accuracy of the algorithm. The calculation formula for indicators is as follows:

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

TP (True Positive) represents the number of positive samples correctly identified; TN (True Negative) represents the number of correctly identified negative samples. FP (False Positive) represents the number of negative samples that were falsely reported. FN (False Negative) represents the number of positive

samples that were missed.

3. Implementation Details

The experiment was conducted on a TPU with 334.6 GB of RAM, utilizing TensorFlow for implementation. Thirty frames were extracted from each video as follows: 10 from the beginning, 10 from the middle, and 10 from the end. The input images were resized to 224 × 224 pixels and normalized. The model architecture was based on NASNetMobile. The Adam optimizer, with its default learning rate of 0.001, was used for training. To prevent overfitting, we froze most layers in the pre-trained NASNetMobile model, allowing only the last five layers to be trainable for fine-tuning. L2 regularization with a factor of 0.001 was applied to the LSTM layers to penalize large weights and improve model generalization. Additionally, dropout layers with a rate of 0.3 were incorporated to randomly deactivate units during training, further enhancing the model's ability to generalize to unseen data.

4. Results

To validate the effectiveness of our proposed model, we compared the latest methods on the Hockey-Fight dataset, Violent-Flows dataset, Real-Life Violence Situations dataset in this section. These methods include Skeleton based methods and Appearance based methods. The comparison results of the three datasets are shown in Table 5 and Table 7, Table 9 respectively.

The experimental results show that the accuracy obtained on the Hockey-Fight dataset is 99.63%. As shown in Table 2, compare with the appearance based methods, including 3D-CNN, I3D,FightNet, Sudhakaran et al., ECO, and TEA.We also compared it with some recent skeleton based methods. Therefore, our method combines these two models and can effectively utilize local and global information. LG-SPIL is using 3D point cloud technology to extract motion feature information of human skeletal points. By introducing the Skeleton Point Interactive Learning (SPIL) module, the model has been improved to 97.5%. The model detection accuracy in this study is 99.63%, which is 2.83% higher than the accuracy of the SPIL. The accuracy of our model is 2.13% higher than the accuracy of the LG-SPIL.

Tableau 5: Comparison with state-of-the-arts on the Hockey-Fight Dataset

Method	Venue(Years)	Video Accuracy(%)
3D-CNN	2014	91.0%
I3D	CVPR2017	93.4%
FightNet	2017	97.0%
Sudhakaran et al.	AVSS2017	97.1%
ECO	ECCV2018	94.0%
TEA	CVPR2020	97.1%
SPIL	ECCV2020	96.8%
Huszár	IEEE2023	97.5%
LG-SPIL	CVPR2023	97.5%
Hachiuma et al	CVPR2023	99.5%
Ours	-	99.63%

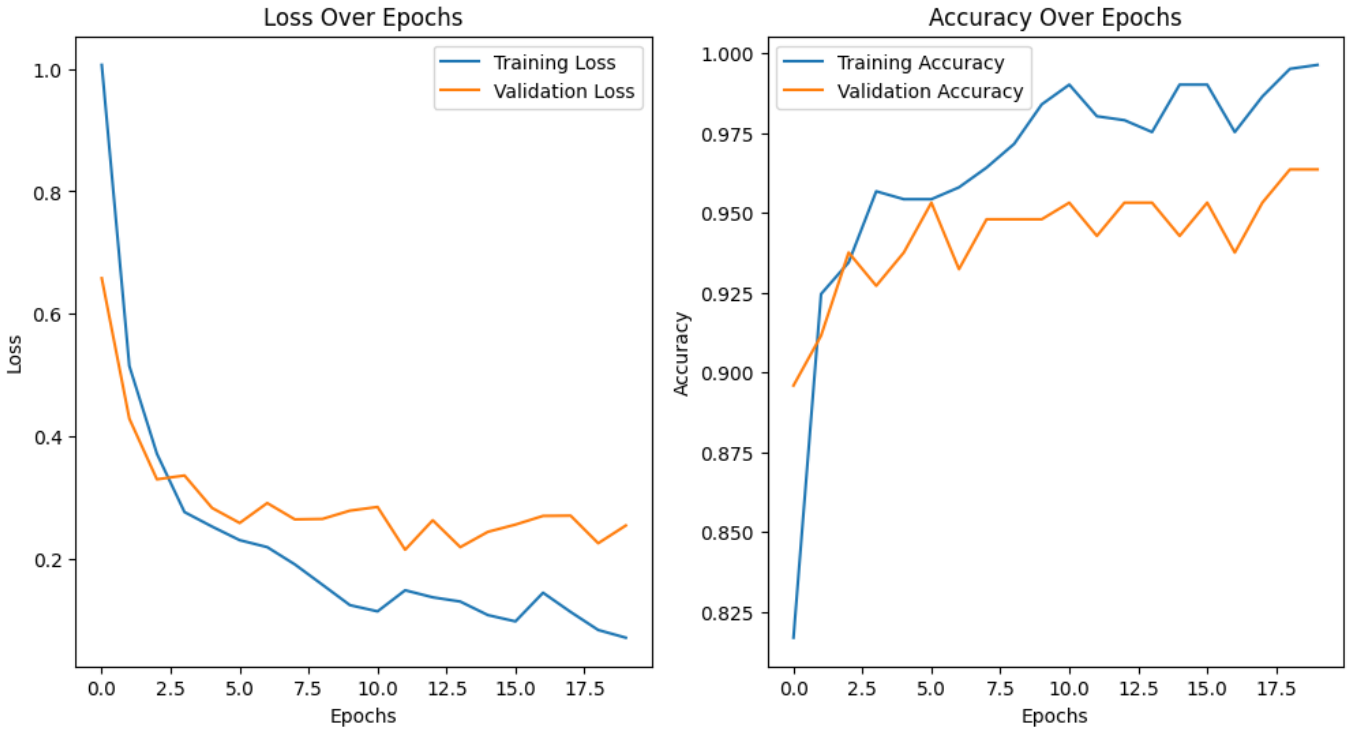


Figure 12: HOCKEY-FIGHT DATASET LOSS AND ACCURACY ON BOTH THE TRAINING AND VALIDATION SETS

Additionally, the confusion matrix in Figure 13 visually presents the distribution of true and predicted classes, reflecting the model's high precision and recall. The detailed classification metrics, including precision, recall, f1-score, and support, are provided in Table 6. These results, summarized below, further confirm the model's robust performance.

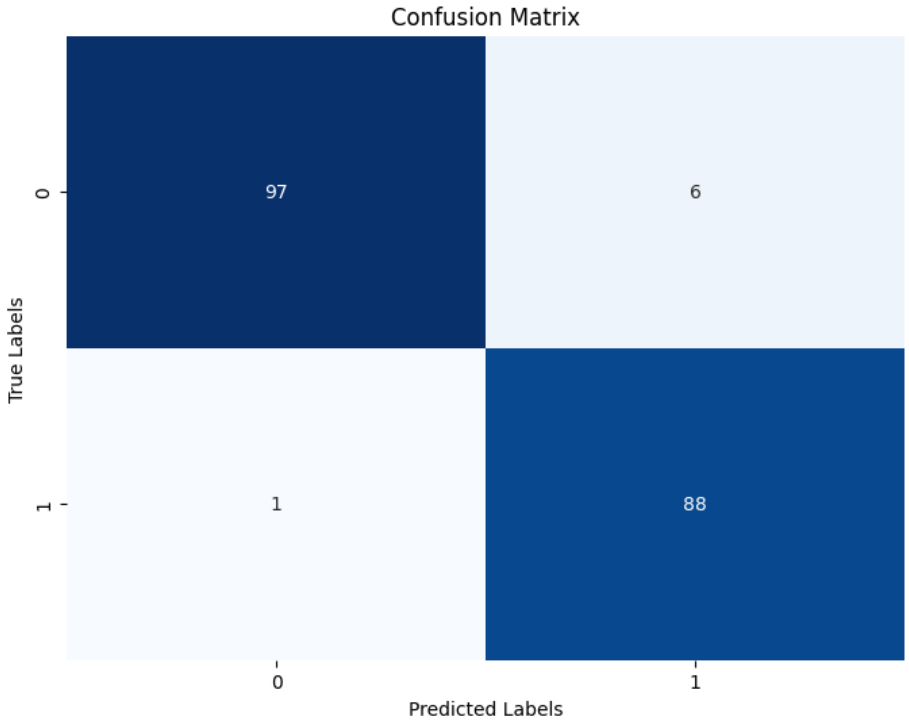


Figure 13: CONFUSION MATRIX FOR THE HOCKEY-FIGHT DATASET

Tableau 6: Classification Metrics for the Hockey-Fight Dataset

Class	Precision	Recall	F1-Score	Support
0	0.99	0.94	0.97	103
1	0.94	0.99	0.96	89

The comparative analysis on the Violent Crowd dataset demonstrates the effectiveness of our proposed method. As shown, our method achieves a perfect accuracy of 100%, outperforming other state-of-the-art techniques, including CNN-BiLSTM (98.64%), 3D CNN (97.0%), and Structured Keypoint Pooling (94.7%). These results highlight the superior performance of our approach in detecting violence within crowd scenes.

Tableau 7: Comparison with state-of-the-arts on the Violent Crowd Dataset

Method	Venue(Years)	Video Accuracy(%)
I3D	CVPR2017	83.4%
SPIL	ECCV2020	95.4%
CNN-BiLSTM	SNComputerScience2020	98.64%
3D CNN	IEEE2022	97.0%
Structured Keypoint Pooling	CVPR2023	94.7%
Proposed	-	**100%**

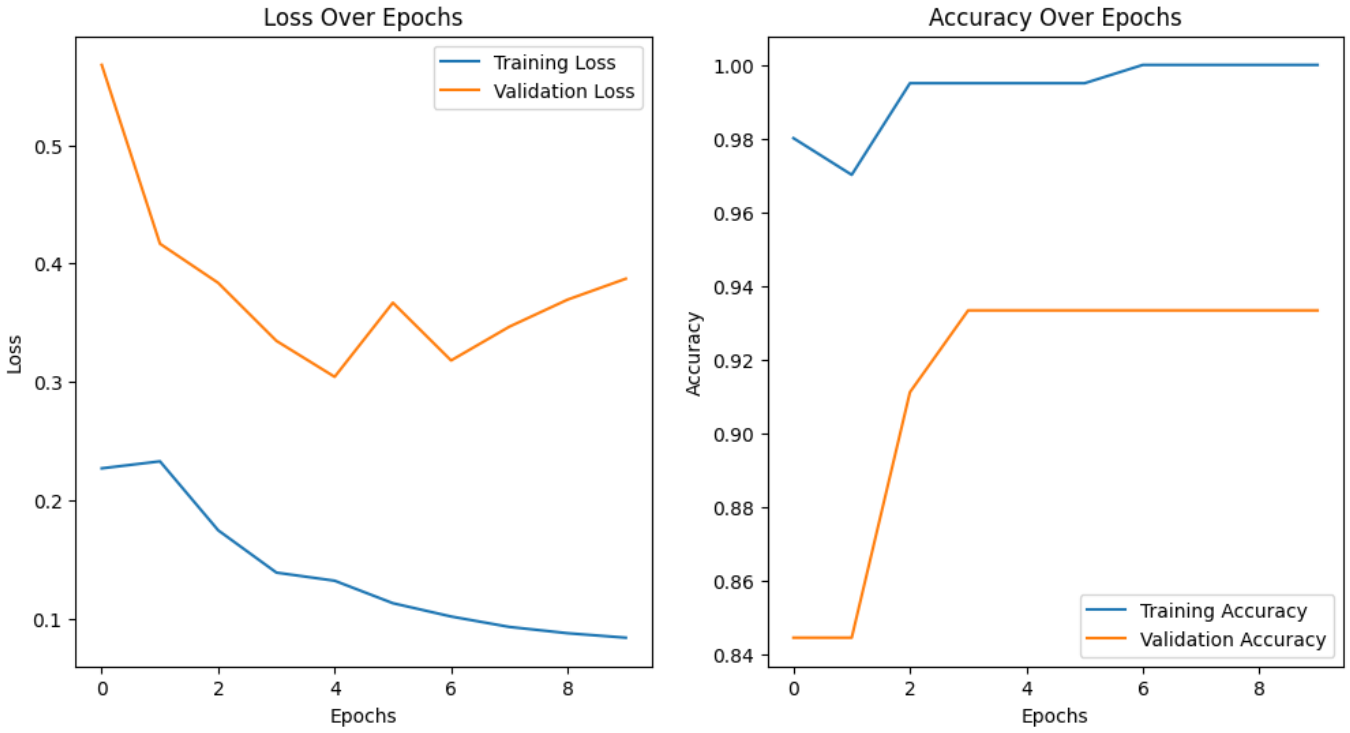


Figure 14: VIOLENT CROWD DATASET LOSS AND ACCURACY ON BOTH THE TRAINING AND VALIDATION SETS

The evaluation of our proposed method on the Violent Crowd dataset is supported by the confusion matrix and the accompanying classification metrics. The confusion matrix highlights the distribution of correct and incorrect predictions, demonstrating the model's proficiency in classifying violent and non-violent scenarios. As detailed in Table 8, the model achieved a precision of 0.96 and a recall of 0.92 for non-violent cases, while for violent cases, it achieved a precision of 0.90 and a recall of 0.95. These metrics, particularly the strong F1-scores for both classes, underscore the model's reliability and effectiveness in detecting violence in complex crowd environments.

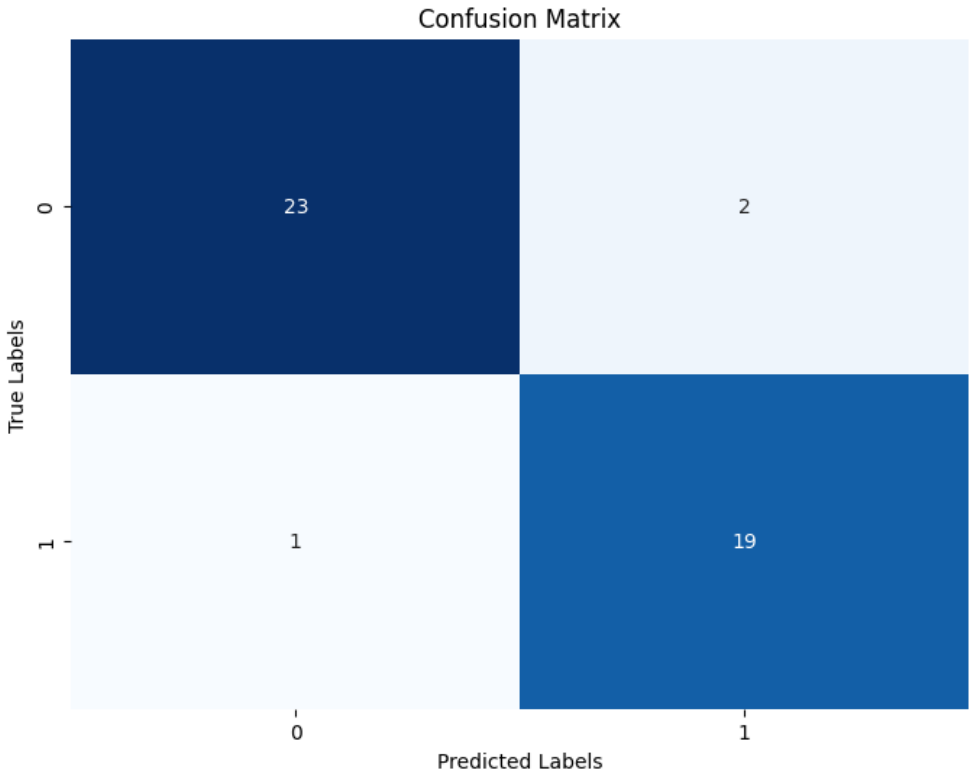


Figure 15: CONFUSION MATRIX FOR THE VIOLENT CROWD DATASET

Tableau 8: Classification Metrics for the Violent Crowd Dataset

Class	Precision	Recall	F1-Score	Support
0	0.96	0.92	0.94	25
1	0.90	0.95	0.93	20

Our proposed method demonstrates also superior performance on the RLSV dataset, achieving an outstanding accuracy of 99.62%. This result surpasses the accuracy of existing methods, underscoring the effectiveness of our approach in accurately detecting violence within various scenes. The high accuracy reflects the model's ability to generalize well across different scenarios in the RLSV dataset, positioning it as a highly reliable solution for violence detection compared to previous techniques.

Tableau 9: Comparison with state-of-the-arts on the RLSV Dataset

Method	Venue(Years)	Video Accuracy(%)
[22]	2021	95.60%
[23]	2022	92.88%
[24]	2023	97.66%
[25]	2023	98.69%
PROPOSED	-	**99.62%**

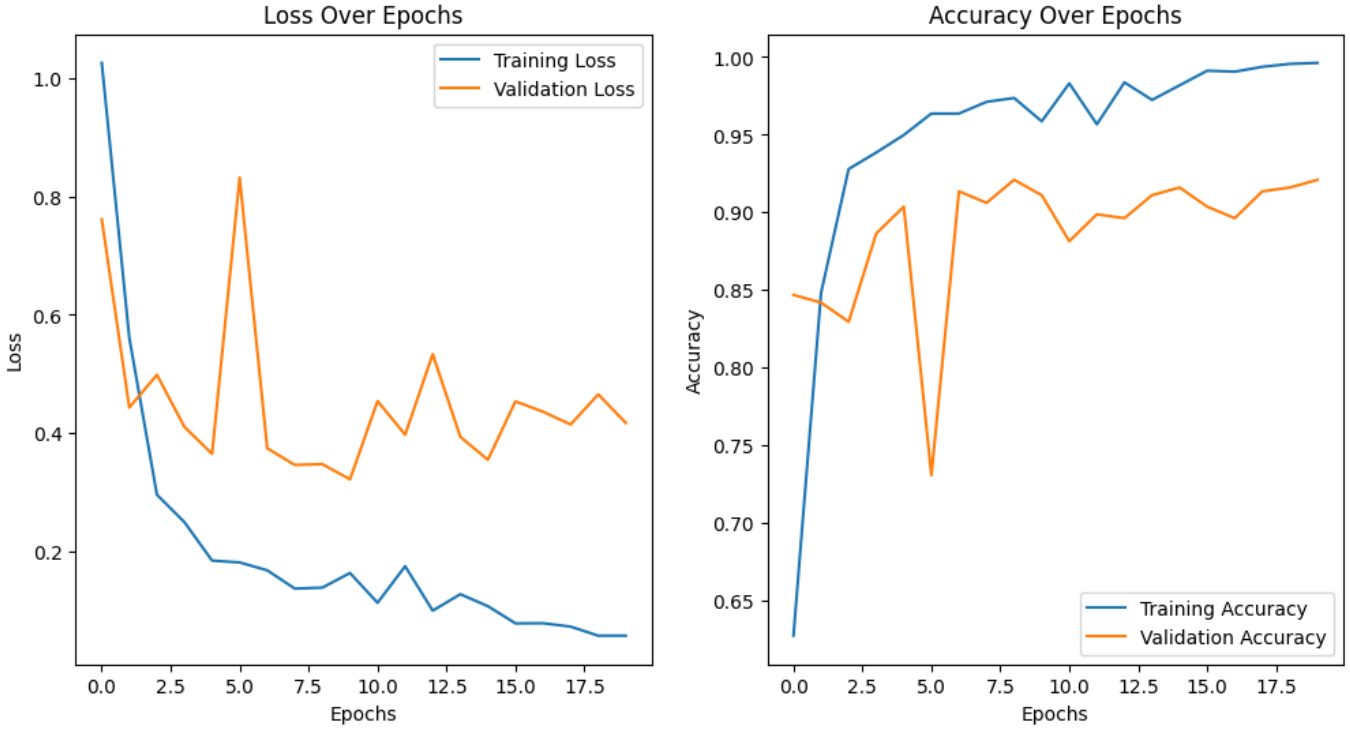


Figure 16: RLSV DATASET LOSS AND ACCURACY ON BOTH THE TRAINING AND VALIDATION SETS

The evaluation of our model on the RLSV dataset is illustrated through the confusion matrix and the classification metrics. The confusion matrix provides insights into the model's accuracy in differentiating between violent and non-violent instances within the dataset. As shown in Table 10, the model achieved a precision of 0.94 and a recall of 0.90 for non-violent cases, and a precision of 0.90 and a recall of 0.95 for violent cases. Both classes have an F1-score of 0.92, reflecting the model's consistent and reliable performance across different types of scenes in the RLSV dataset.

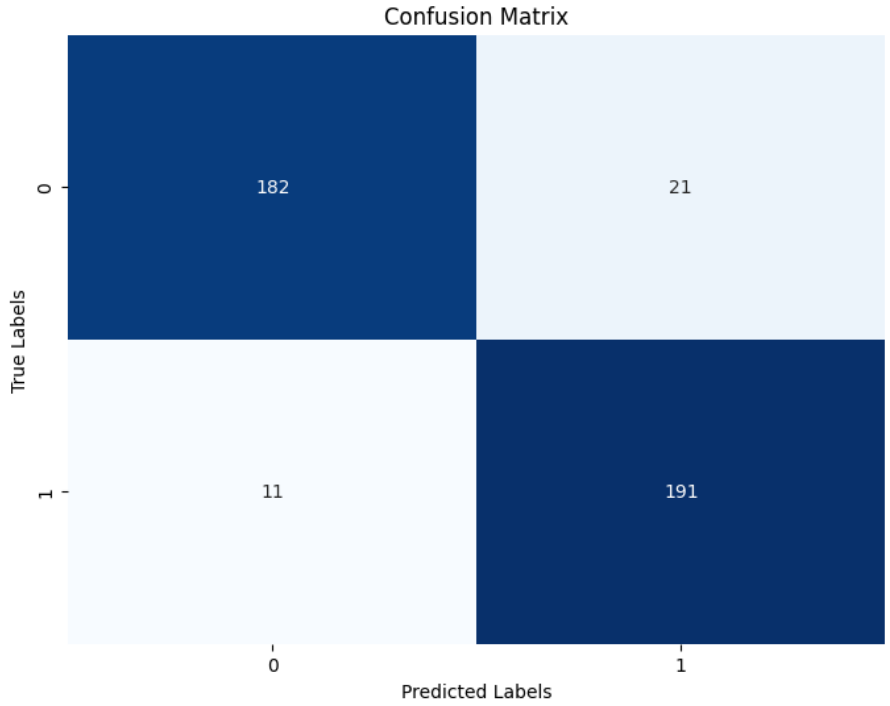


Figure 17: CONFUSION MATRIX FOR THE RLSV DATASET

Tableau 10: Classification Metrics for the RLSV Dataset

Class	Precision	Recall	F1-Score	Support
0	0.94	0.90	0.92	203
1	0.90	0.95	0.92	202

5. Discussion and limitaions

The integration of NASNetMobile, LSTM, and an attention mechanism in this two-stream framework shows great promise in accurately detecting violent behavior by effectively combining RGB and pose data; however, the reliance on precise pose estimation and the computational intensity of the approach present challenges that must be addressed for practical deployment in real-world scenarios; continued refinement of the model, particularly in terms of robustness and efficiency, is essential for ensuring its applicability across a wider range of environments and conditions.

The framework's effectiveness heavily depends on the accuracy of pose estimation, which can be compromised in scenarios involving occlusion, low resolution, or unusual body positions, leading to reduced detection performance; the integration of NASNetMobile and LSTM, while powerful, imposes significant computational demands that may hinder deployment on resource-constrained devices such as mobile phones or edge devices; additionally, the model's ability to generalize across diverse real-world environments, including varying lighting conditions, backgrounds, and camera angles, remains a challenge that could affect its robustness in less controlled settings.

6. Future Work

Future research could delve deeper into enhancing pose estimation accuracy and robustness, especially in challenging scenarios like crowded environments, occlusions, or low-light conditions, by integrating more sophisticated models or employing refined data fusion techniques. For example, advancements in multi-scale networks, graph convolutional networks (GCNs), or transformers could be leveraged to improve the understanding of human poses in complex settings. Additionally, incorporating temporal dynamics through techniques such as spatiotemporal modeling could enable the system to more accurately track and predict human movements over time, further refining violence detection capabilities.

Beyond improving pose estimation, exploring model optimization strategies presents another critical area for future research. Techniques such as model pruning, quantization, or knowledge distillation could be utilized to reduce the computational overhead and memory usage, making these systems more efficient for real-time deployment. The adoption of lightweight architectures, such as MobileNets or efficient deep learning models, could allow violence detection systems to be implemented on edge devices or in low-resource settings without sacrificing accuracy. This could open new possibilities for integrating violence detection technology into portable devices or in scenarios with limited processing power, like drones or smart surveillance cameras.

Furthermore, expanding the current frameworks to include additional modalities, such as audio signals or contextual scene information, could significantly boost detection accuracy and robustness. Audio-based cues, like raised voices or sudden loud noises, could complement visual data to provide a more holistic understanding of potentially violent situations. Integrating contextual scene information, such as location data or object detection, could help systems infer the broader context of a scene, allowing for better differentiation between violent and non-violent interactions in varied environments. Multimodal systems that combine visual, auditory, and contextual inputs could therefore offer more comprehensive violence detection solutions.

Another promising direction for future research involves broader validation on diverse and more complex datasets. Validating these models on datasets that include a wide range of violent behaviors, cultural contexts, and complex environmental factors—such as weather conditions, lighting variations, and background clutter—would help improve the generalizability and robustness of the model. This would ensure that violence detection systems can perform reliably across different regions, scenarios, and forms of aggression, from subtle altercations to more overt physical violence. Researchers could also explore adversarial testing methods to ensure the system’s robustness against intentional manipulation, such as camouflage or deceptive behavior designed to bypass detection.

Additionally, there could be significant value in exploring unsupervised or semi-supervised learning techniques for improving model performance in cases where labeled data is scarce. By leveraging unlabeled data or utilizing few-shot learning methods, future research could address the data limitation challenges that often arise in violence detection projects. This would enable the model to adapt and learn from new or previously unseen forms of violent behavior without requiring extensive manual annotation efforts.

Ultimately, a combination of advancements in pose estimation, model optimization, multimodal integration, and validation on diverse datasets will be essential for pushing the boundaries of violence detection technology. These innovations could bring about more robust, efficient, and contextually aware systems that play a critical role in public safety, surveillance, and real-time intervention.

Conculsion

This report presents a real-time violence detection framework based on a two-stream approach that leverages both RGB and pose estimation data. Our method employs NASNetMobile as the backbone for feature extraction, LSTM for temporal modeling, and an attention mechanism to selectively focus on the most relevant features. This framework enhances violence detection by integrating both visual appearance and pose information, allowing the model to effectively capture and interpret complex human movements. By using this dual-stream architecture and attention mechanism, the model adaptively focuses on either RGB or pose data depending on the context, improving its robustness in diverse scenarios.

To validate our approach, extensive experiments were conducted on multiple video-level benchmarks. The proposed method achieved state-of-the-art accuracy, with 99.63% on the Hockey-Fight dataset, 100% on the Violent Crowd dataset, and 99.62% on the RLSV dataset. These results demonstrate the effectiveness of combining RGB, pose information, and attention mechanisms in enhancing violence detection performance, significantly surpassing existing methods.

List of Figures

List of tables

References

[1] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, X. He, A new method for violence detection in surveillance scenes, Multimedia Tools Appl. 75 (12) (2016) 7327–7349.
[2] P. Zhou, Q. Ding, H. Luo, X. Hou, Violence detection in surveillance video using low-level features, PLoS One 13 (10) (2018) e0203668.
[3] A.S. Saif, Z.R. Mahayuddin, Moment features based violence action detection using optical flow, Int. J. Adv. Comput. Sci. Appl. 11 (11) (2020) 503–510.
[4] Q. Xu, J. See, W. Lin, Localization guided fight action detection in surveillance videos, in: 2019 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2019, pp. 568–573.
[5] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan, et al., Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM, Comput. Netw. 151 (2019) 191–200.
[6] P. Zhou, Q. Ding, H. Luo, X. Hou, Violent interaction detection in video based on deep learning, J. Phys.: Conf. Ser. 844 (2017) 012044.
[7] Irfanullah, T. Hussain, A. Iqbal, B. Yang, A. Hussain, Real time violence de tection in surveillance videos using convolutional neural networks, Multimedia Tools Appl. 81 (26) (2022) 38151–38173.
[8] F. De Souza, H. Pedrini, Detection of violent events in video sequences based on census transform histogram, in: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI, IEEE, 2017, pp. 323–329.
[9] J. Mahmoodi, A. Salajeghe, A classification method based on optical flow for violence detection, Expert Syst. Appl. 127 (2019) 121–127.
[10] P.C. Ribeiro, R. Audigier, Q.C. Pham, RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance, Comput. Vis. Image Underst. 144 (2016) 121–143.
[11] K. Lloyd, P.L. Rosin, D. Marshall, S.C. Moore, Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures, Mach. Vis. Appl. 28 (3) (2017) 361–371.
[12] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, Z. Zheng, MoWLD: A robust motion image descriptor for violence detection, Multimedia Tools Appl. 76 (1) (2017) 1419–1438.
[13] K. Deepak, L. Vignesh, S. Chandrakala, Autocorrelation of gradients based violence detection in surveillance videos, ICT Express 6 (3) (2020) 155–159.
[14] O. Deniz, I. Serrano, G. Bueno, T.-K. Kim, Fast violence detection in video, in: 2014 International Conference on Computer Vision Theory and Applications, Vol. 2, VISAPP, IEEE, 2014, pp. 478–485.
[15] S. Mukherjee, R. Saini, P. Kumar, P.P. Roy, D.P. Dogra, B.-G. Kim, Fight detection in hockey videos using deep network, J. Multimedia Inf. Syst. 4 (4) (2017) 225–232.
[16] F.A. Pujol, H. Mora, M.L. Pertegal, A soft computing approach to violence detection in social media for smart cities, Soft Comput. 24 (15) (2020) 11007–11017.
[17] T. Senst, V. Eiselein, T. Sikora, A local feature based on Lagrangian measures for violent video classification, in: 6th International Conference on Imaging for Crime Prevention and Detection, ICDP-15, IET, 2015, pp. 1–6.
[18] T. Senst, V. Eiselein, A. Kuhn, T. Sikora, Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation, IEEE Trans. Inf. Forensics Secur. 12 (12) (2017) 2945–2956.
[19] E.Y. Fu, H.V. Leong, G. Ngai, S. Chan, Automatic fight detection based on motion analysis, in:

2015 IEEE International Symposium on Multimedia, ISM, 2015, pp. 57–60.

[20] Talha, K.R.; Bandapadya, K.; Khan, M.M. Violence Detection Using Computer Vision Approaches. In Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 544–550.
[21] Madhavan,R.; Utkarsh.; Vidhya, J. Violence Detection from CCTV Footage Using Optical Flow and Deep Learning in Inconsistent Weather and Lighting Conditions. In Proceedings of the Advances in Computing and Data Sciences: 5th International Conference, ICACDS2021, Nashik, India, 23–24 April 2021; Revised Selected Papers, Part I 5. Springer: Berlin/Heidelberg, Germany, 2021; pp. 638–647.

[22] F.J. Rendón-Segador, J.A. Álvarez-García, F. Enríquez, O. Deniz, Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence, Electronics 10 (13) (2021) 1601.
[23] D. Choqueluque-Roman, G. Camara-Chavez, Weakly supervised violence detection in surveillance video, Sensors 22 (12) (2022) 4502.
[24] S.A. Jebur, K.A. Hussein, H.K. Hoomod, L. Alzubaidi, Novel deep feature fusion framework for multi-scenario violence detection, Computers 12 (9) (2023) 175.
[25] C. Li, X. Yang, G. Liang, Keyframe-guided video swin transformer with multi-path excitation for violence detection, Comput. J. (2023) bxad103.

[26] Magdy, M.; Fakhr, M.W.; Maghraby, F.A. Violence 4D: Violence detection in surveillance using 4D convolutional neural networks. *IET Computer Vision* 2023, 17, 282–294.

[27] Chen, Y.; Zhang, B.; Liu, Y. ESTN: Exacter Spatiotemporal Networks for Violent Action Recognition. In *Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, Nanjing, China, 22–24 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 44–48.

[28] Wintarti, A.; Puspitasari, R.D.I.; Imah, E.M. Violent Videos Classification Using Wavelet and Support Vector Machine. In *Proceedings of the 2022 International Conference on ICT for Smart Society (ICISS)*, Bandung, Indonesia, 10–11 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 01–05.

[29] Lohithashva, B.; Aradhya, V.M. Violent video event detection: A local optimal oriented pattern-based approach. In *Proceedings of the Applied Intelligence and Informatics: First International Conference, AII 2021*, Nottingham, UK, 30–31 July 2021; Proceedings 1. Springer: Berlin/Heidelberg, Germany, 2021; pp. 268–280.

[30] Ullah, F.U.M.; Obaidat, M.S.; Muhammad, K.; Ullah, A.; Baik, S.W.; Cuzzolin, F.; Rodrigues, J.J.; de Albuquerque, V.H.C. An intelligent system for complex violence pattern analysis and detection. *Int. J. Intell. Syst.* 2022, 37, 10400–10422.

[31] Vijeikis, R.; Raudonis, V.; Dervinis, G. Efficient violence detection in surveillance. *Sensors* 2022, 22, 2216.

[32] Halder, R.; Chatterjee, R. CNN-BiLSTM model for violence detection in smart surveillance. *SN Comput. Sci.* 2020, 1, 201.

[33] Traoré, A.; Akhloufi, M.A. 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. In *Proceedings of the International Conference on Image Analysis and Recognition*, Póvoa de Varzim, Portugal, 24–26 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 152–160.

[34] Aarthy, K.; Nithya, A.A. Crowd Violence Detection in Videos Using Deep Learning Architecture. In *Proceedings of the 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 16–17 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.

[35] Asad, M.; Yang, J.; He, J.; Shamsolmoali, P.; He, X. Multi-frame feature-fusion-based model for violence detection. *Vis. Comput.* 2021, 37, 1415–1431.

[36] Contardo, P.; Tomassini, S.; Falcionelli, N.; Dragoni, A.F.; Sernani, P. Combining a mobile deep neural network and a recurrent layer for violence detection in videos. In *Proceedings of the RTA-CSIT 2023: 5th International Conference Recent Trends and Applications in Computer Science and Information Technology*, Tirana, Albania, 26–27 April 2023.

[37] Gupta, H.; Ali, S.T. Violence Detection using Deep Learning Techniques. In *Proceedings of the 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, Hyderabad, India, 25–27 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 121–124.

[38] Islam, M.S.; Hasan, M.M.; Abdullah, S.; Akbar, J.U.M.; Arafat, N.; Murad, S.A. A deep Spatio-temporal network for vision-based sexual harassment detection. In *Proceedings of the 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, Dhaka, Bangladesh, 21–23 December 2021; IEEE: Piscataway, NJ, USA; 2021, pp. 1–6.

[39] Jahlan, H.M.B.; Elrefaei, L.A. Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video. *Arab. J. Sci. Eng.* 2021, 46, 8549–8563.

[40] Mumtaz, N.; Ejaz, N.; Aladhadh, S.; Habib, S.; Lee, M.Y. Deep multi-scale features fusion for effective violence detection and control charts visualization. *Sensors* 2022, 22, 9383.

[41] Sharma, S.; Sudharsan, B.; Naraharisetti, S.; Trehan, V.; Jayavel, K. A fully integrated violence detection system using CNN and LSTM. *Int. J. Electr. Comput. Eng.* (2088-8708) 2021, 11, 3374–3380.

[42] Singh, N.; Prasad, O.; Sujithra, T. Deep Learning-Based Violence Detection from Videos. In *Proceedings of the Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, Mizoram, India, 25–26 June 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 323–332.

[43] Srivastava, A.; Badal, T.; Saxena, P.; Vidyarthi, A.; Singh, R. UAV surveillance for violence detection and individual identification. *Autom. Softw. Eng.* 2022, 29, 28.

[44] Ullah, F.U.M.; Obaidat, M.S.; Muhammad, K.; Ullah, A.; Baik, S.W.; Cuzzolin, F.; Rodrigues, J.J.; de Albuquerque, V.H.C. An intelligent system for complex violence pattern analysis and detection. *Int. J. Intell. Syst.* 2022, *37*, 10400–10422.
[45] Vijeikis, R.; Raudonis, V.; Dervinis, G. Efficient violence detection in surveillance. *Sensors* 2022, *22*, 2216.

[46] Halder, R.; Chatterjee, R. CNN-BiLSTM model for violence detection in smart surveillance. *SN Comput. Sci.* 2020, *1*, 201.
[47] Traoré, A.; Akhloufi, M.A. 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. In *Proceedings of the International Conference on Image Analysis and Recognition*, Póvoa de Varzim, Portugal, 24–26 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 152–160.
[48] Aarthy, K.; Nithya, A.A. Crowd Violence Detection in Videos Using Deep Learning Architecture. In *Proceedings of the 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 16-17 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
[49] Asad, M.; Yang, J.; He, J.; Shamsolmoali, P.; He, X. Multi-frame feature-fusion-based model for violence detection. *Vis. Comput.* 2021, *37*, 1415–1431.
[50] Contardo, P.; Tomassini, S.; Falcionelli, N.; Dragoni, A.F.; Sernani, P. Combining a mobile deep neural network and a recurrent layer for violence detection in videos. In *Proceedings of the RTA-CSIT 2023: 5th International Conference Recent Trends and Applications in Computer Science and Information Technology*, Tirana, Albania, 26–27 April 2023.
[51] Gupta, H.; Ali, S.T. Violence Detection using Deep Learning Techniques. In *Proceedings of the 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, Hyderabad, India, 25–27 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 121–124.
[52] Islam, M.S.; Hasan, M.M.; Abdullah, S.; Akbar, J.U.M.; Arafat, N.; Murad, S.A. A deep Spatio-temporal network for vision-based sexual harassment detection. In *Proceedings of the 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, Dhaka, Bangladesh, 21–23 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
[53] Jahlan, H.M.B.; Elrefaei, L.A. Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video. *Arab. J. Sci. Eng.* 2021, *46*, 8549–8563.
[54] Mumtaz, N.; Ejaz, N.; Aladhadh, S.; Habib, S.; Lee, M.Y. Deep multi-scale features fusion for effective violence detection and control charts visualization. *Sensors* 2022, *22*, 9383.
[55] Sharma, S.; Sudharsan, B.; Naraharisetti, S.; Trehan, V.; Jayavel, K. A fully integrated violence detection system using CNN and LSTM. *Int. J. Electr. Comput. Eng.* (2088-8708) 2021, *11*, 3374–3380.
[56] Singh, N.; Prasad, O.; Sujithra, T. Deep Learning-Based Violence Detection from Videos. In *Proceedings of the Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, Mizoram, India, 25–26 June 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 323–332.
[57] Srivastava, A.; Badal, T.; Saxena, P.; Vidyarthi, A.; Singh, R. UAV surveillance for violence detection and individual identification. *Autom. Softw. Eng.* 2022, *29*, 28.
[58] Islam, Z.; Rukonuzzaman, M.; Ahmed, R.; Kabir, M.H.; Farazi, M. Efficient two-stream network for violence detection using separable convolutional LSTM. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, Virtual, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
[59] Mahmoodi, J.; Nezamabadi-pour, H.; Abbasi-Moghadam, D. Violence detection in videos using interest frame extraction and 3D convolutional neural network. *Multimed. Tools Appl.* 2022, *81*, 20945–20961.
[60] Ahmed, M.; Ramzan, M.; Khan, H.U.; Iqbal, S.; Khan, M.A.; Choi, J.I.; Nam, Y.; Kadry, S. Real-Time Violent Action Recognition Using Key Frames Extraction and Deep Learning; Tech Science Press: Henderson, NV, USA, 2021.
[61] Ji, Y.; Wang, Y.; Kato, J.; Mori, K. Predicting Violence Rating Based on Pairwise Comparison.

IEICE Trans. Inf. Syst. 2020, *103*, 2578–2589.
[62] Ehsan, T.Z.; Mohtavipour, S.M. Vi-Net: A deep violent flow network for violence detection in video sequences. In *Proceedings of the 2020 11th International Conference on Information and Knowledge Technology (IKT)*, Tehran, Iran, 22–23 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 88–92.
[63] Jayasimhan, A.; Pabitha, P. A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset. In *Proceedings of the 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4)*, Bangalore, India, 15–16 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
[64] Kim, H.; Jeon, H.; Kim, D.; Kim, J. Lightweight framework for the violence and falling-down event occurrence detection for surveillance videos. In *Proceedings of the 2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Republic of Korea, 19–21 October 2022; IEEE: Piscataway, NJ, USA; 2022; pp. 1629–1634.
[65] Monteiro, C.; Durães, D. Modelling a Framework to Obtain Violence Detection with Spatial-Temporal Action Localization. In *Proceedings of the World Conference on Information Systems and Technologies*, Galicia, Spain, 16–19 April 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 630–639.
[66] Zhang, Z.; Yuan, D.; Li, X.; Su, S. Violent Target Detection Based on Improved YOLO Network. In *Proceedings of the International Conference on Artificial Intelligence and Security*, Qinghai, China, 15–20 July 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 480–492.
[67] Appavu, N. Violence Detection Based on Multisource Deep CNN with Handcraft Features. In *Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, Hammamet, Tunisia, 29 April–1 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
[68] Adithya, H.; Lekhashree, H.; Raghuram, S. Violence Detection in Drone Surveillance Videos. In *Proceedings of the International Conference on Smart Computing and Communication*, Nashville, TN, USA, 26–30 June 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 703–713.
[69] Bi, Y.; Li, D.; Luo, Y. Combining keyframes and image classification for violent behavior recognition. *Appl.Sci.*2022,*12*,8014.
[70] Freire-Obregón, D.; Barra, P.; Castrillón-Santana, M.; Marsico, M.D.80. Qu, W.; Zhu, T.; Liu, J.; Li, J. A time sequence location method of long video violence based on improved C3D network. J. Supercomput. 2022, 78, 19545–19565.
[71]. Zhou, L. End-to-end video violence detection with transformer. In Proceedings of the 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 19–21 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 880–884.
[72]. Hung, L.P.; Yang, C.W.; Lee, L.H.; Chen, C.L. Constructing a Violence Recognition Technique for Elderly Patients with Lower Limb Disability. In Proceedings of the International Conference on Smart Grid and Internet of Things; Springer: Cham, Germany, 2021; pp. 24–37.
[73]. Mahalle, M.D.; Rojatkar, D.V. Audio based violent scene detection using extreme learning machine algorithm. In Proceedings of the 2021 6th international conference for convergence in technology (I2CT), Maharashtra, India, 2–4 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
[74]. Zheng, Z.; Zhong, W.; Ye, L.; Fang, L.; Zhang, Q. Violent scene detection of film videos based on multi-task learning of temporal-spatial features. In Proceedings of the 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), Tokyo, Japan, 22–24 March 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 360–365.
[75]. Aktı, S.; Ofli, F.; Imran, M.; Ekenel, H.K. Fight detection from still images in the wild. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI,

USA, 3–8 January 2022; pp. 550–559.
[76]. Ehsan, T.Z.; Nahvi, M.; Mohtavipour, S.M. An accurate violence detection framework using unsupervised spatial–temporal action translation network. Vis. Comput. 2023, 40, 1515–1535.
[77]. Ullah, F.U.M.; Muhammad, K.; Haq, I.U.; Khan, N.; Heidari, A.A.; Baik, S.W.; de Albuquerque, V.H.C. AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. IEEE Trans. Ind. Inform. 2021, 18, 5359–5370.
[78]. Santos, F.; Durães, D.; Marcondes, F.S.; Lange, S.; Machado, J.; Novais, P. Efficient violence detection using transfer learning. In Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems, Salamanca, Spain, 6–8 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 65–75.
[79]. Sernani, P.; Falcionelli, N.; Tomassini, S.; Contardo, P.; Dragoni, A.F. Deep learning for automatic violence detection: Tests on the AIRTLab dataset. IEEE Access 2021, 9, 160580–160595.
[80]. Shang, Y.; Wu, X.; Liu, R. Multimodal Violent Video Recognition Based on Mutual Distillation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Shenzhen, China, 4–7 November 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 623–637.
[81]. Hua, G.; Li, L.; Liu, S. Multipath affinage stacked—Hourglass networks for human pose estimation. Front. Comput. Sci. 2020, 14, 1–12.
[82]. Liu, S.; Li, Y.; Hua, G. Human pose estimation in video via structured space learning and halfway temporal evaluation. IEEE Trans. Circuits Syst. Video Technol. 2018, 29, 2029–2038.
[83]. Mohtavipour, S.M.; Saeidi, M.; Arabsorkhi, A. A multi-stream CNN for deep violence detection in video sequences using handcrafted features. Vis. Comput. 2022, 38, 2057–2072.
[84]. Jaiswal, S.G.; Mohod, S.W. Classification of Violent Videos Using Ensemble Boosting Machine Learning Approach with Low Level Features. Indian J. Comput. Sci. Eng. 2021, 12, 1789–1802.
[85]. Hu, X.; Fan, Z.; Jiang, L.; Xu, J.; Li, G.; Chen, W.; Zeng, X.; Yang, G.; Zhang, D. TOP-ALCM: A novel video analysis method for violence detection in crowded scenes. Inf. Sci. 2022, 606, 313–327.
[86]. Naik, A.J.; Gopalakrishna, M. Deep-violence: Individual person violent activity detection in video. Multimed. Tools Appl. 2021, 80, 18365–18380.
[87]. Narynov, S.; Zhumanov, Z.; Gumar, A.; Khassanova, M.; Omarov, B. Detecting School Violence Using Artificial Intelligence to Interpret Surveillance Video Sequences. In Proceedings of the Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, 29 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 401–412.
[88]. Srivastava, A.; Badal, T.; Garg, A.; Vidyarthi, A.; Singh, R. Recognizing human violent action using drone surveillance within real-time proximity. J. Real-Time Image Process. 2021, 18, 1851–1863.
[89]. Su, Y.; Lin, G.; Zhu, J.; Wu, Q. Human interaction learning on 3d skeleton point clouds for video violence recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 74–90.
[90]. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 322–339.
[91]. Cheng, S.T.; Hsu, C.W.; Horng, G.J.; Jiang, C.R. Video reasoning for conflict events through feature extraction. J. Supercomput. 2021, 77, 6435–6455.
[92]. Kumar, A.; Shetty, A.; Sagar, A.; Charushree, A.; Kanwal, P. Indoor Violence Detection using Lightweight Transformer Model. In Proceedings of the 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 26–28 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.

[93] Contardo, P.; Tomassini, S.; Falcionelli, N.; Dragoni, A.F.; Sernani, P. Combining a mobile deep neural network and a recurrent layer for violence detection in videos. In Proceedings of the RTA-CSIT 2023: 5th International Conference Recent Trends and Applications in Computer Science and Information Technology, Tirana, Albania, 26–27 April 2023.