China Instrument and Control Society
Instrumentation
INSTRUMENTATION, Vol. 11, No. 4, December 2024 https://doi.org/10.15878/j.instr.202400230

Article

A violence detection method based on deep and shallow feature fusion

Lin'en Liu, Xuguang Zhang*

HangZhou Dianzi University, The Communication Engineering Department, Hangzhou 310018, China

* Corresponding author email: zhangxg@hdu.edu.cn

Abstract: In the research of video-based violent behavior detection, the motion information in the video is vital for violence detection. How to highlight motion information in videos and integrate spatiotemporal information is an urgent problem that needs to be solved in violence detection. In this paper, we propose a deep learning architecture that integrates shallow features into deep features to strengthen the network's ability to express motion information at a deep level. To enhance the weight of motion information in the network, we design a downsampling module to extract shallow features, fused with the deep features extracted by MobileNet's Blocks. Furthermore we constructed a channel attention module and introduced a Convolutional Long Short-Term Memory (ConvLSTM) module. These two modules aim to redistribute network attention: the channel attention module focuses on channel-level information and the ConvLSTM module emphasizes temporal aspects. Finally, we employ 3D convolution and global pooling to compress the feature sizes, fed into fully connected layers to perform violence detection. Experiments are conducted on three publicly available standard datasets, achieving an accuracy rate of 91% on the surveillance video dataset RWF2000, 97.5% on the Hockey fight dataset, and 100% on the movies dataset. Overall, the proposed model demonstrates satisfactory performance in violence detection.

Keywords: surveillance; optical flow; violence detection; deep learning

Citation: Lin'en Liu, Xuguang Zhang."A violence detection method based on deep and shallow feature fusion." Instrumentation 11, no.4 (December 2024). https://doi.org/10.15878/j.instr.202400230

1 Introduction

In recent years, surveillance cameras have become integral public infrastructure, and the development and application of surveillance video systems have positively impacted social security[1]. Built upon this foundation, intelligent security systems play a crucial role in advancing smart cities, owing to their capability to enhance urban security management through crowd behavior detection[2]. The automatic recognition of abnormal behaviors or events in videos has been widely applied in numerous fields such as surveillance and public safety[3]. As the number of surveillance videos recording human activities rises, there is now a growing demand for the automatic identification of violent events that jeopardize social security. As a result, violence detection has emerged as a prominent research focus in computer vision[4].

Neural networks have made significant contributions in various fields, including their application to violence detection. In the era of big data, the growing abundance of videos containing violence offers substantial and diverse databases to train deep learning models. Firstly, two-dimensional (2D) convolutional neural networks (CNNs), known for their excellent performance in image recognition tasks, have been applied to violence detection. However, they face challenges in understanding video data[5]. Videos depicting violent behavior are essentially temporal sequences, and relying solely on human limb gestures captured in a single image lacks the contextual dynamic motion information needed to identify violent behavior accurately. Recurrent neural

networks (RNNs) are adept at handling temporal sequences, as demonstrated in natural language processing[6]. Introducing RNNs can assist CNNs in learning temporal features effectively. The Long Short-Term Memory (LSTM) architecture is commonly used for processing temporal sequences. When endowed with convolutional structures, LSTM evolves into ConvLSTM, enabling the network to encode temporal information across frames while extracting spatial features.

Due to hardware limitations, surveillance devices face challenges in handling neural networks with large parameters. We tend to search for a network model with fewer parameters while achieving higher recognition accuracy. The MobileNet series performs excellently in large-scale classification tasks, utilizing significantly fewer parameters than large neural networks. This makes it well-suited for deployment on surveillance devices for real-time recognition tasks[7]. Due to the limited number of samples in publicly available violence recognition datasets, it is difficult to train a new neural network from scratch. Therefore, in this paper, a pre-trained MobileNet V3 model is utilized. We extracted the optical flow and grayscale frame differences from the video sequence and concatenated them to form the input, instead of using the original input image. The new input can directly reflect the motion information of violent behavior, enhancing its suitability for violence detection tasks. However, the splicing input introduces an additional challenge. The pre-trained MobileNet model does not incorporate optical flow or frame difference images as part of its pre-training process. As a result, the deeper features of the network tend to be biased toward recognizing static objects rather than capturing dynamic motion information. Building upon the concept of the residual structure[8], motion features that remain figuratively expressed in the shallow layer are merged into the deeper abstract features following several downsampling stages. A multi-scale sampling module is constructed using three parallel convolutional layers with different kernel sizes to better utilize shallow features. This module is combined with ReLU, Batch Normalization and Pooling layers to form a downsampling module. The features obtained through downsampling are merged with the deep features of the network using the proposed fusion strategy. This merging augments the significance of motion features within the network, thereby enhancing its comprehension and expression of dynamic information. An attention mechanism is incorporated at the end of the feature fusion module to emphasize the deeper features. Convolutional layers are employed to reduce the number of channels in the fused features while achieving information correlation among the channels. Global pooling maps each channel's feature map into a corresponding weight, which is then computed by a convolutional layer and multiplied with the feature map, effectively redistributing the feature weights across channels. Then, the ConvLSTM module is utilized to process the entire video sequence, where the features outputted by the hidden layer at each time step establish temporal correlations with the entire video.

The main contributions of our proposed model include:

1. We designed a downsampling module to extract shallow features of the neural network and then fused them with the deep features to enhance the weight of motion information within the network.

2. We designed a channel attention module to emphasize the channel weights of the feature maps for each frame. The ConvLSTM module was also introduced to strengthen the temporal correlation between fragmented individual frame features and integrate them into a unified whole.

In the remainder of this paper, a comprehensive review of the existing literature on neural networks for violence detection is provided in Section 2. Section 3 elaborates on the image preprocessing method, the network model architecture, and our proposed module. Section 4 presents the experimental results obtained from our methodology. Finally, a summary of our work is provided in Section 5.

2 Related work

Numerous publicly available CNN models have exhibited excellent performance in image recognition, forming a progressive series of neural network models. These models have been trained on extensive image classification datasets that are publicly available. They show promising potential to transfer to other tasks, significantly reducing the time required to learn fundamental features. The progressive evolution and refinement of the image domain in the realm of deep learning have significantly broadened the applicability of recognition methodologies, enabling neural networks to handle complex human behavior that can effectively understand the intricacies and dynamics of human interactions within a video[9]. In violence detection research, which often centers on video analysis, pre-trained CNN models hold significant potential for transfer learning. This capability enables the identification of violent behavior within target videos.

When dealing with a video sequence composed of multiple frames, it is imperative to consider the recognition results of the entire video sequence to identify violent behavior. The 2D CNN utilizes its convolutional layers to extract spatial features from video frames, resulting in a one-dimensional (1D) feature vector as its output. These vectors from the same video are concatenated and fed into the recurrent neural network for learning temporal features. 2D CNNs play a pivotal role in acquiring the spatial features of the images. Kumar et al. [9] used a pre-trained Inception v3 model, Ullah et al. [10] utilized a pre-trained ResNet-50

model, and Sarcar et al. [11] employed a pre-trained VGG16 model for their respective studies. These studies demonstrate the effectiveness of using established CNN architectures for violence detection tasks. First, the spatial features of all frames are extracted and connected back to a unified sequence. Then, temporal dependency is captured by an LSTM module, and subsequently, the classification results are obtained through the fully connected layer. At the final convolutional layer, the feature maps, before being converted into 1D vectors, contain more information, and the temporal information of the feature map sequence can be extracted by a ConvLSTM module. The spatial features of video frames are extracted by the convolutional layers of the VGG16[12] or VGG19[13]. Next, the ConvLSTM module replaces the fully connected layers of VGG to incorporate the temporal information across video frames, establishing temporal correlations. Finally, new fully connected layers are appended to the ConvLSTM module to obtain the ultimate classification results.

The 2D CNN mainly relies on convolutional kernels to extract spatial features. In contrast, the 3D convolutional kernel enhances this capability by adding a temporal dimension to the 2D convolutional kernel, enabling the extraction of both spatial and temporal features. Therefore, the 3D CNN can directly learn the whole video sequence. Most of the research conducted on 3D CNNs revolves around the innovation and development of the C3D model[14][15]. The Two-Stream Inflated 3D ConvNets (I3D) [16] model matches the temporal length of the video by inflating the convolutional kernel. The two-stream input design of RGB and optical flow provides more feature information for the network[17]. Inflating the 2D convolutional kernel enables I3D to utilize parameters from pre-trained 2D CNNs. The subsequent advancement of I3D marks a new milestone and starting point in 3D CNN research, influencing subsequent studies[18]. Furthermore, 3D CNNs can serve as a complement to 2D CNNs, concatenating the feature maps learned by both networks in the channel dimension can provide richer information[19].

Among the prevailing CNNs, 2D CNNs are primarily designed for single-image recognition tasks, which restricts their ability to capture comprehensive temporal features. Even with supplementation by LSTM, they still face challenges in effectively learning complete spatio-temporal features. On the other hand, 3D CNNs have the inherent capability to learn spatio-temporal features of video frames directly. However, their usage often entails higher computational complexity and more parameters, presenting challenges regarding computational resources and training efficiency. When the real-time recognition capability is considered an essential factor in evaluating model performance, existing 2D CNNs offer a broader array of options. This paper adopts the MobileNet as the backbone network to reduce parameters and enhance computational efficiency. Furthermore, the ConvLSTM was employed to capture temporal correlations. The improved model based on these foundations demonstrates increased recognition accuracy.

3 The proposed methodology

In our proposed end-to-end neural network model for detecting violent behavior in videos, our methodology involves leveraging 2D CNNs to capture spatial features from each input frame. A channel attention module was utilized to reconstruct the channel weights of the feature maps, and then a ConvLSTM module was applied to integrate the video sequence along the temporal dimension. These extracted features were forwarded to a classifier constructed with fully connected layers, ultimately facilitating the determination of whether violent behavior is present.

3.1 Optical flow and frame difference extraction

The main evidence used to determine violence in the video objectively focuses on the behavioral actions of the crowd, often disregarding complex background information beyond the moving subject. By stripping the moving subject from the video and eliminating the interference from the background, the attention of the neural network can be directed toward the actions of the crowd in the video. However, raw images only depict the spatial relationship between characters and do not convey their movement process, limiting the neural network to learning static information from such inputs. By using optical flow, the motion information within the video can be captured, revealing how the limb positions of the characters change over time and highlighting the instantaneous velocity of the moving subject. We employed the Gunnar Farneback optical flow method[20] to estimate the displacement of light for every pixel between two consecutive frames, which provides values representing the horizontal and vertical directions of movement. Among the various methods for calculating optical flow, the Farneback optical flow method stands out for its sensitivity to abnormal scenes[21] and offers distinct advantages in detecting abnormal crowds. Additionally, subtracting pixel values between consecutive frames can eliminate the image background, while the frame difference method preserves the contour features of the moving subject. These approaches capture some temporal information while protecting the primary spatial information. Both optical flow and frame difference are suitable for input in violence detection tasks.

The utilization of two-stream inputs offers a more comprehensive representation of information, thereby enhancing the recognition capability of the model. However, the computational demands are approximately doubled compared to a single-stream network. Moreover, the three color channels of the raw image often

encompass similar information, leading to a degree of redundancy. The optical flow from two channels and the frame difference represented in grayscale are merged into a unified input using a concatenation approach. This approach aimed to emulate the input effect seen in two-stream neural networks. In contrast to the raw image, which primarily conveys color information, the splicing input offers a more precise representation of motion information. In Equation 1, *Frame* is the original color image of the video and $Frame_t$ is the *tth* frame in the video sequence.

$$Input_t = Cat[Flow_{(x,y)}(Frame_{t+1}, Frame_t), Gray(|Frame_{t+1} - Frame_t|)] \quad (1)$$

3.2 Network architecture

The sample size of existing violence detection datasets is insufficient to meet the order of magnitude required for deep learning. Therefore, it is suitable to use the method of transfer learning to fine-tune the neural network trained on a large dataset. To balance the computational cost and recognition accuracy, we chose to use MobileNet to extract spatial features. MobileNet V3[22] adds the Squeeze-and-Excite module to the Inverted Residuals Block, constructing several similar Linear Bottlenecks structures responsible for primary feature extraction tasks. We extracted the shallow features before the Blocks, introduced them into the parallel downsampling module and then concatenated them with the deep features. The concatenated features underwent convolutional fusion to align channel dimensions, followed by multiplication with the deep features and then addition with them, as shown in Figure 1. The fused feature maps were passed through a channel attention module to update the channel weights. Following the passage of all the input video frames through the CNN, they were merged into a single unified video sequence. Subsequently, a layer of the ConvLSTM module[23] was used to establish the temporal dependency. A 3D convolution was applied to compress the temporal dimension, followed by global average pooling to reduce the size of the feature map. Ultimately, the four-dimensional (4D) tensor, which includes dimensions for time, channel, height, and width, was transformed into a 1D tensor. This sensor had a length equal to the number of channels and served as the input to the fully connected layer to detect violent behavior in the video.

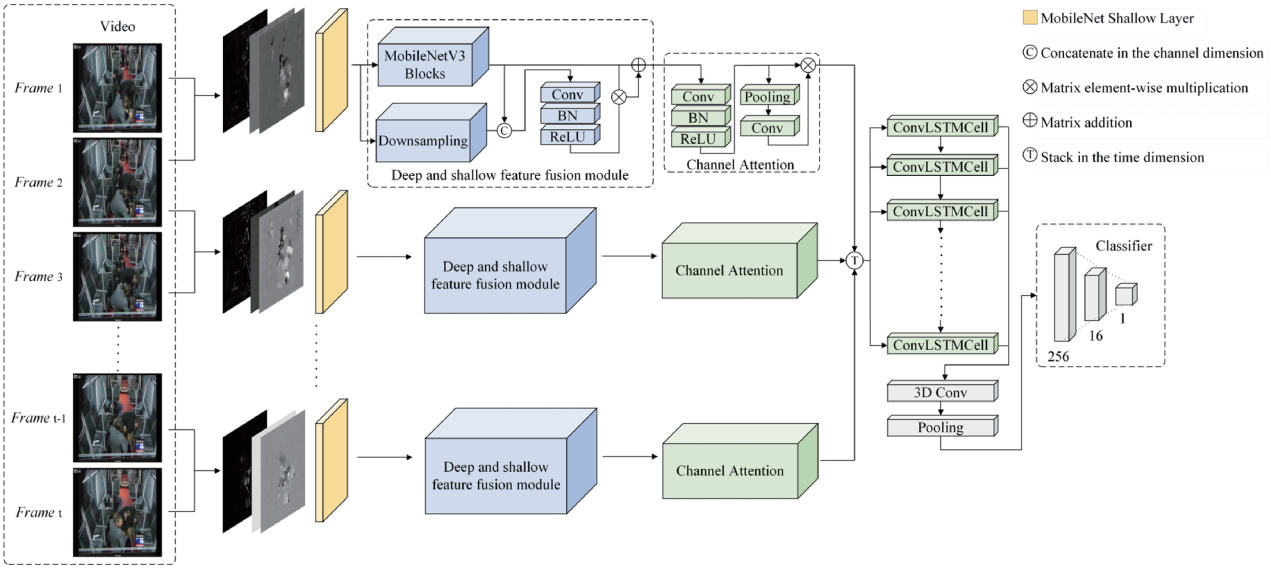


Fig.1 The Overall flowchart of the network and the fusion module and attention mechanism

3.3 Deep and shallow feature fusion

The pre-trained MobileNet provides an interface for customizing the classifier, meaning that the model allows users to train end-of-line classifier modules on their datasets. Therefore, we removed the fully connected layer for classification and retained the convolutional layer-based Blocks. Inserting a new module into a pre-trained neural network renders the network parameters beyond that module obsolete. Therefore, the optimal strategy is to merge the features obtained from downsampling into the last layer of the Blocks. The features, obtained through downsampling modules after only a few feature extraction steps, exhibit less semantic information compared to the deep features extracted by the Blocks. They cannot independently perform violence detection and merely complement the deep features.

The shallow features still show more concrete representational information. We designed appropriate convolutional layers within the downsampling module to extract the abstract violent characteristics. Due to the diverse camera angles and distances used during video recording, the proportion of moving subjects in the images differs significantly. A single receptive field is insufficient to comprehensively extract information from subjects of different sizes. To effectively utilize shallow motion information across various scales, we proposed a

multi-scale feature extraction module, as shown in Figure 2. This module employs a parallel three-branch structure, where each branch utilizes convolution kernels of three distinct sizes for downsampling. In the Blocks, feature extraction and channel expansion are primarily conducted using a 3×3 convolutional kernel. Consequently, we designed the 3x3 convolutional layer to generate new feature maps with C/2 channels, supplemented by 1×1 and 5×5 convolutional kernels in separate layers to produce feature maps with C/4 channels. These feature maps were concatenated to get the downsampled feature maps with C channels.

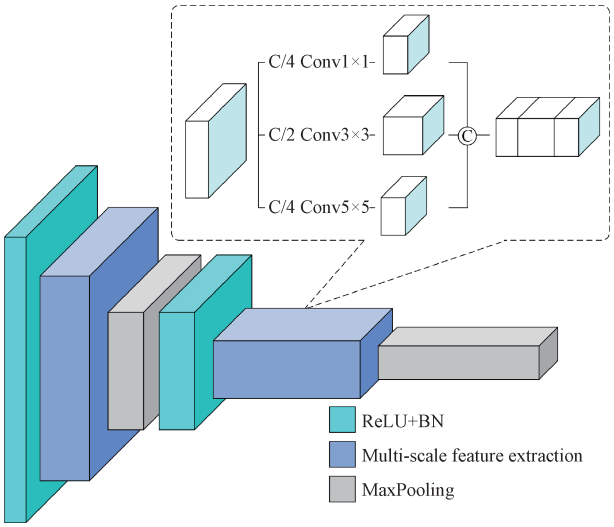


Fig.2 The layer structure of the downsampling module

The ReLU activation function, Batch Normalization (BN), a multi-scale feature extraction module, and MaxPooling are used to construct the downsampling module, as shown in Figure 2. The activation function enabled the neural network to learn more complex nonlinear relationships, thus increasing the network's generalization ability. BN adjusted the data to a standard normal distribution, alleviating gradient vanishing problems and enhancing network stability. ReLU and BN were used with the convolutional layers to mitigate the risk of overfitting. MaxPooling merged the local regions of the feature map through a sliding pooling window, which reduced the size of the feature map while retaining the most salient feature values. This process helped to suppress extraneous noise and redundant information. The multi-scale feature extraction module extracted shallow motion features while expanding the number of channels, thereby creating more feature space to capture more intricate patterns. The sampling step size of the three convolution kernels in the module was set to 2 to downsize the feature map. With MaxPooling, the module accomplished the downsampling of the shallow features, reducing the number of parameters and computational complexity in the model. Using the aforementioned module, the downsampling module running in parallel with the Blocks unified the feature map size of the shallow features before the Blocks and the deep features after them, enabling seamless feature fusion in subsequent stages.

Table 1 presents the specific parameters of the two proposed modules within the deep and shallow feature fusion module. These parameters include the sizes of the input and output feature maps at each layer, the kernel sizes in convolutional and pooling layers, as well as the strides. Convolutional layers with the same suffix notation are arranged in a parallel structure and belong to the same multi-scale feature extraction module.

Table 1 Parameters of each layer in the deep and shallow feature fusion module

Input	Operator	Kernel size	Stride	Output
Downsampling module layers				
16×112^2	ReLU	/	/	16×112^2
16×112^2	BatchNorm	/	/	16×112^2
16×112^2	**Conv1_1**	1	2	20×56^2
16×112^2	**Conv3_1**	3	2	40×56^2
16×112^2	**Conv5_1**	5	2	20×56^2
80×56^2	MaxPool	2	2	80×28^2
80×28^2	ReLU	/	/	80×28^2
80×28^2	BatchNorm	/	/	80×28^2
80×28^2	**Conv1_2**	1	2	240×14^2
80×28^2	**Conv3_2**	3	2	480×14^2
80×28^2	**Conv5_2**	5	2	240×14^2
960×14^2	MaxPool	2	2	960×7^2
Fusion module layers				
1920×7^2	Conv	1	1	960×7^2
960×7^2	ReLU	/	/	960×7^2
960×7^2	BatchNorm	/	/	960×7^2

In Equation 2, the deep features extracted by the Blocks and the shallow features obtained by downsampling in the channel dimension are concatenated using a convolutional layer and adjusting the number of channels. Subsequently, BN and ReLU activation functions are applied to strengthen the nonlinear representation of the model.

$$feature_{fusion} = ReLU(BN(Conv(Cat[feature_{blocks}, feature_{down}]))) \quad (2)$$

In Equation 3, the fused features are multiplied by the deep features and then added to the latter to enhance common important features while attenuating less significant features and suppressing noise.

$$feature'_{blocks} = feature_{blocks} + feature_{fusion} \times feature_{blocks} \quad (3)$$

At this stage, the MobileNet model and the deep and shallow feature fusion module have completed spatial

feature extraction and enhanced motion features for each frame of the input.

3.4 Channel attention and temporal correlation

We devised a channel attention module to leverage the fused features effectively. The fused feature map undergoes a convolutional operation to decrease the number of channels. Following this reduction, in Equation 4, analogous to the ECA (Efficient Channel Attention) module [24], the feature maps $feature^{C\times W\times H}$ undergo global pooling, allowing each channel's feature map to be represented by its corresponding value. Distinct from employing convolutional kernels of adaptive sizes to correlate feature information between channels, the proposed module achieves channel interdependency simultaneously through the preceding channel reduction before pooling. Consequently, the convolutional kernel size here is set to 1, enabling the weight values $W^{C\times 1\times 1}$ to focus more on their corresponding channels.

$$W^{C\times 1\times 1} = Conv(Pooling(feature^{C\times W\times H})) \tag{4}$$

In Equation 5, the weights $W^{C\times 1\times 1}$ are multiplied by the feature maps $feature^{C\times W\times H}$ to complete the redistribution of the channel weights.

$$feature_{new}^{C\times W\times H} = feature^{C\times W\times H} \times W^{C\times 1\times 1} \tag{5}$$

After the CNN with the channel attention module, each frame is converted to feature vectors with dimensions C×W×H. Stacking the feature vectors corresponding to each frame in chronological order to obtain a 4D feature vector of T×C×W×H, where T represents the temporal dimension. However, simple stacking can only be seen as the process where multiple static images collectively contribute to the recognition of violent behavior, disregarding the contextual relationships among consecutive frames. To enable the model to learn the dynamic process of violent behavior, the ConvLSTM was introduced to understand the temporal correlation of the video sequences. As shown in Figure 3, the vectors are input into ConvLSTM in a temporal order. Cell State($C_1, C_2, \cdots, C_t$) is the memory unit of ConvLSTM, preserving the previous information and contributing to the computation of the new information. Based on the previous Cell State and Hidden State, along with the current vector V_t, the output at the current time can be calculated, which is the Hidden State H_t. Stacking the Hidden States of the ConvLSTM Cell outputs in chronological order yields the $Output^{T\times C\times W\times H}$ of the ConvLSTM, which is the real information encoded by the ConvLSTM. Each Hidden State is strongly related to the current moment and can establish a temporal relationship with the inputs of all previous moments, which aids the model in adaptively assigning time weights to video sequences. Hence, we utilized $Output^{T\times C\times W\times H}$ as the ultimate feature vector. This vector effectively captures the temporal dynamics and encodes the contextual relationships among consecutive frames, enhancing the model's ability to recognize violent behavior.

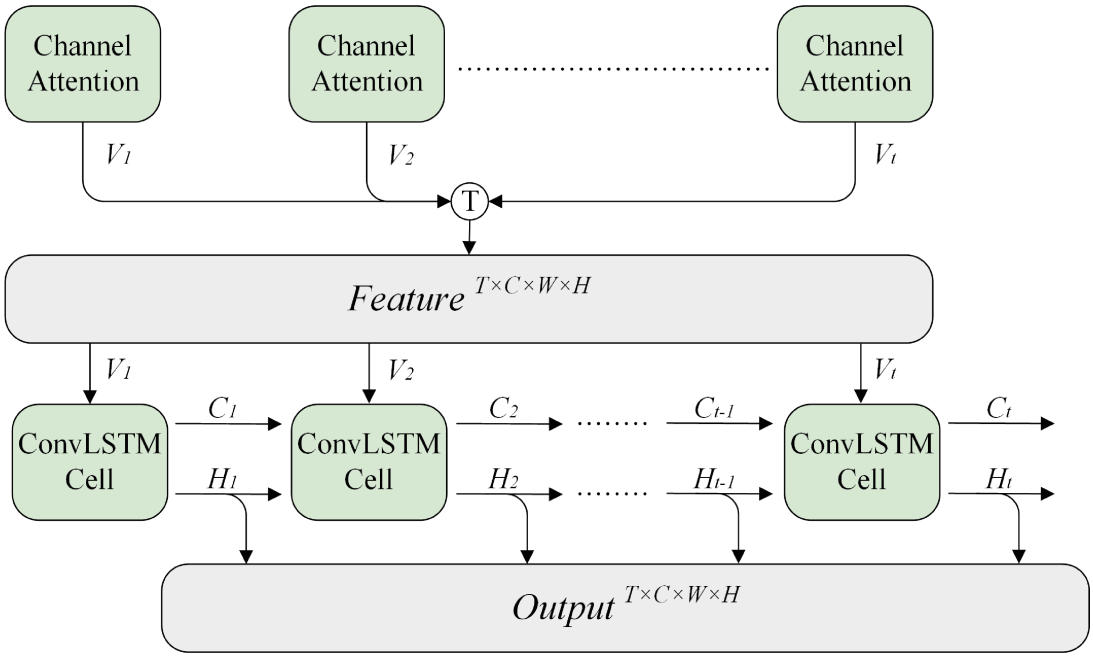


Fig.3 The ConvLSTM establishes temporal dependencies among isolated vectors.

4 Experimental results and analysis

In this section, we detail our training methodology, including the programming language utilized in our experiments and the necessary hyperparameters for training. Following this, we introduce the publicly

available dataset used in this paper. Finally, the experimental results were analyzed from multiple perspectives. Our proposed approach demonstrated excellent recognition capability in our comparison of recognition accuracy across various methods on the same dataset. We conducted ablation experiments to assess the impact of our proposed method on model performance. Heatmaps were generated for different positions within the deep and shallow feature fusion module using samples from the validation set, effectively demonstrating the functionality of this module. Additionally, we constructed confusion matrices and performed detailed analysis on representative samples to showcase the performance of our proposed model under specific conditions. The results indicated that the improved neural network is better suited for violence detection.

4.1 Hyperparameter settings

The video duration varied across different datasets. To standardize the length of the video sequence, average sampling was employed to extract 30 frames to construct the input video sequence. Each frame tensor was sized at 30×3×320×320. Our proposed method was implemented using Python 3.9 and the PyTorch. During the training process, we set up 100 epochs to train the model. Several common image enhancement techniques enriched the data samples, including random cropping, random horizontal flipping, brightness and hue adjustment, and data normalization. The optimizer was AdamW, with the initial learning rate set to 0.00001 and the weight decay to 0.00002. A cosine annealing strategy was employed to reduce the learning rate during the training process dynamically. The loss function was binary cross entropy (BCE). For the MobileNetV3 model, we selected the Large version and retained all pre-trained parameters. As for other network modules, default parameters were used for their initial settings.

4.2 Datasets information

The Hockey Fight dataset[25] comprises 1,000 videos captured during hockey games, where a mobile camera tracks the movements of hockey players. The videos vary in length, with violent video clips primarily capturing instances of fighting among the players during the game. In contrast, nonviolent video clips showcase the regular progression of the game. Notably, the recorded behavioral actions in this dataset exhibit a relative similarity, with backgrounds consistently depicting sports venues dedicated to ice hockey games.

The Movies dataset[25] consists of video clips sourced from various movies, comprising 100 violent clips and 100 nonviolent clips. This dataset is typically presented within the same article as the Hockey Fight dataset and it shares similar clip lengths to the latter. Unlike the Hockey Fight dataset that only has a single background type, video clips in the Movies dataset contain more diverse background types.

The RWF2000 dataset[26] consists of 2000 surveillance video clips that have been pre-divided by the author into a training set and a validation set. The training set comprises 1600 clips, and the validation set shall consist of 400 video clips. Each video clip in this dataset has a fixed duration of 5 seconds, but some videos cannot play at normal speeds. These video clips are captured by surveillance cameras and encompass a range of real-life scenes. The videos often contain factors such as obstructed characters and low shooting light, which are highly relevant to violence detection applications.

4.3 Network performance comparison

In the dataset used for our experiment, 80% of the video samples were used to train our model, and the remaining portion was used to test its performance. As shown in Table 2, our model achieves a notable level of accuracy on the RWF2000 dataset, surpassing that of most other network models. In video samples captured by surveillance cameras, the position of objects and the brightness of light remain largely unchanged in the background, while the moving subject can be clearly highlighted in optical flow and frame difference. Our proposed model demonstrates excellent performance on datasets captured with fixed cameras. Samples from the Hockey Fight dataset are videos captured by a moving camera, where changes in the relative position of characters cannot be accurately mapped to their actual movements. As a result, our proposed model exhibits slightly lower accuracy compared to other methods on this particular dataset. As evaluated on the Movies dataset across different backgrounds, our proposed model demonstrates near-perfect recognition.

The parameters of the proposed model are presented in Table 3. Compared to the approach using 3D CNNs, the proposed model boasts a smaller number of parameters and has lower requirements for hardware platforms. During the experiments, a GTX 1080Ti was sufficient to complete the entire training process. A test sample can be detected within 0.4 seconds.

4.4 Ablation experiment

We conducted ablation experiments on the RWF2000 and Hockey Fight dataset, evaluating all combinations of splicing input and two proposed modules. Table 4 demonstrates the improvement achieved by our proposed method on the model's recognition performance. It can be seen that the table, the splicing input plays a significant role in improving the performance of the model. In addition, enabling the 2D CNN to directly learn the motion features facilitates the detection of behavioral actions. The deep and shallow feature fusion module enhances the pre-trained network's adaptability to violence detection. It achieves this by reusing the shallow feature information to strengthen the weight of the training samples in the network. The module also facilitates the transfer of the abstract

Table 2 Comparison of accuracy with other methods on standard datasets

Method	Hockey Fight	Movies	RWF2000
ViF[27]	82.9%	\	\
HOG+Random Forest[28]	86%	\	\
Inception-Resnet-V2[29]	93.33%	100%	\
ResNet50+NN[30]	96%	100%	\
U-Net+ResNet50[31]	96.4%	\	\
C3D+Inception-Resnet-V2[32]	95.1%	\	\
MobileNetV2+SepConvLSTM, 2 Streams[33]	99%	100%	89.75%
3D CNN+3D CA+ConvLSTM[34]	\	\	89.7%
Yolov3+ConvLSTM+GRU[35]	98.5%	\	88.2%
HRNet+3D CNN[36]	\	\	89.45%
3D CNN[37]	95.5%	100%	73%
I3D, 2 Streams[38]	98.7%	99%	90.4%
ConvLSTM[39]	94.5%	98.5%	90.25%
VST+GCN+VGG+BiLSTM[40]	97.97%	\	90.74%
ResNet50+POT[41]	96%	100%	84%
X3D[42]	\	\	87.2%
Ours	97.5%	100%	91%

Table 3 Comparison of accuracy and parameters on the RWF2000 dataset

Method	Accuracy	Parameters(M)
HRNet+3D CNN[36]	89.45%	13.54
I3D, 2 Streams[38]	90.4%	13.2
Ours	91%	6.52

representation of the network's deep features from visual object recognition to violence detection. The channel attention module and ConvLSTM adaptively learn important features from different dimensions. Together, they form an attention mechanism that allows the model to focus more accurately on information related to violent behavior. However, due to the low difficulty of the Hockey Fight dataset, all methods have good effects on it. According to the ablation experiment, the proposed method has slightly improved performance, but there is not much difference between various methods. At the same time. Using the raw image input yields better results than combining it with optical flow. This is because the cameras in the Hockey Fight dataset are in motion, so inputting the original image is better. However, considering both of the two datasets, the method of integrating optical flow has stronger adaptability.

Table 4 Comparison of proposed methods on model performance

Input	Model	RWF2000	Hockey Fight
Raw image input	MobileNet	78.25%	97%
Raw image input	MobileNet+Fusion	81.25%	96.5%
Raw image input	MobileNet+Attention	80.5%	96%
Raw image input	MobileNet+Fusion+Attention	82%	98%
Splicing input	MobileNet	84.75%	96%
Splicing input	MobileNet+Fusion	88.5%	95%
Splicing input	MobileNet+Attention	86%	96%
Splicing input	MobileNet+Fusion+Attention	91%	97.5%

4.5 Visualized heat map

We used the trained model to detect violent samples in the validation set and generated its heat map through the deep and shallow feature fusion module. The detection process of the network can be visualized through the heat map, as shown in Figure 4. The first column shows the raw images in the samples. In the second column, the heat map of the shallow layer reveals activated thermal regions that are relatively scattered. Moving to the third column, the heat map represents the downsampling module, while the fourth column illustrates the heat map of the Blocks. The heat maps of violent samples demonstrate that the downsampling module and the Blocks accurately focus on the areas where violent behavior occurs. The fifth column displays the heat map of the fusion module, highlighting a region of activation that is more concentrated and accurate. This module seamlessly combines the deep and shallow features, leading to enhanced accuracy in localizing violent behavior. The sixth column presents the heatmap of MobileNet without the proposed module. When compared to the heatmap exhibited by the fusion module in the fifth column, the unmodified network demonstrates inferior sensitivity to violent actions and accuracy in the region of interest. The comparison of the heatmap between the unmodified MobileNet and the fusion module is consistent with the quantitative accuracy comparison results from the ablation experiments. This underscores the proposed method's effectiveness in distinguishing between violent and non-violent scenes by actively attending to violence information and relevant motion patterns.

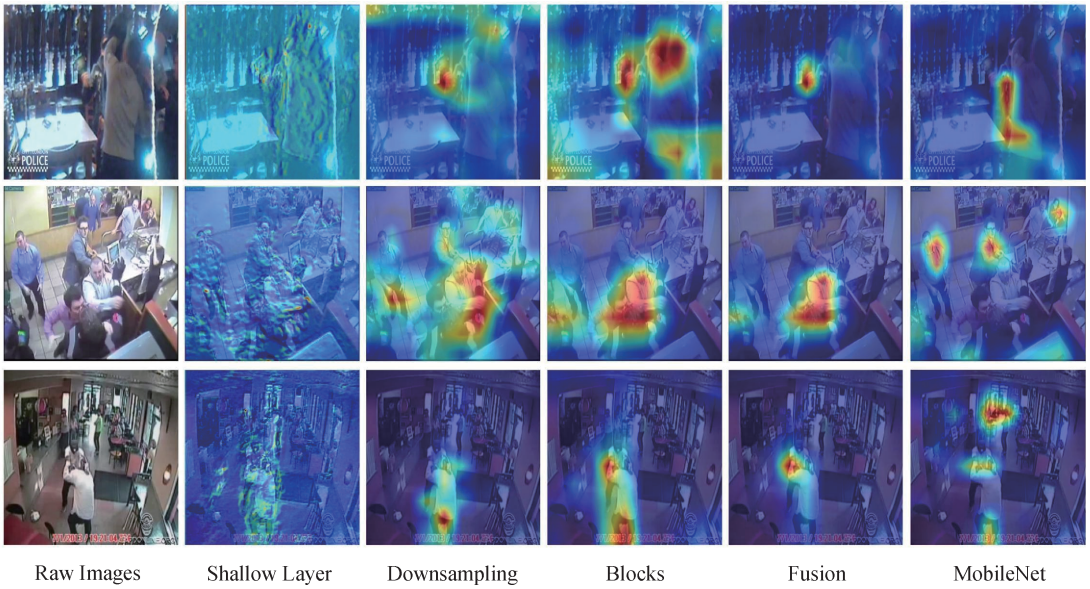


Fig.4 The heat map generated by the test sample passing through different parts of the networks

4.6 Confusion matrix analysis

We considered samples containing instances of violence as Positive and those without violence as Negative. Our proposed model was then applied to test the datasets, generating the confusion matrices, as shown in Figure 5.

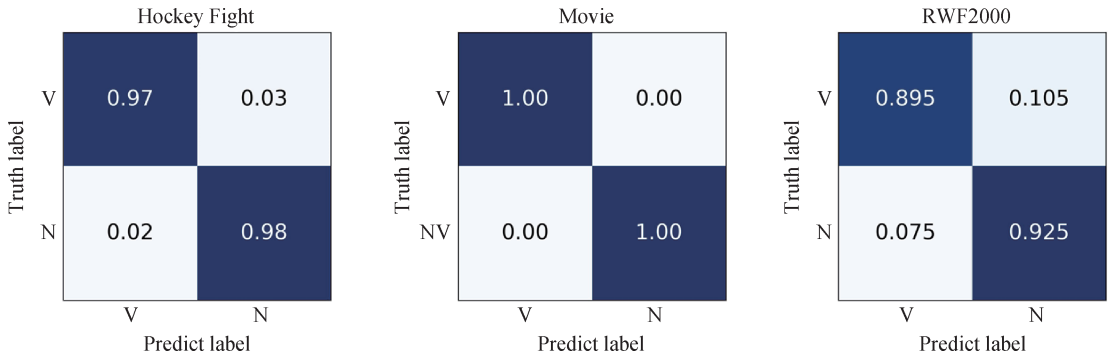


Fig.5 Confusion matrices depicting the accuracy obtained by the proposed model

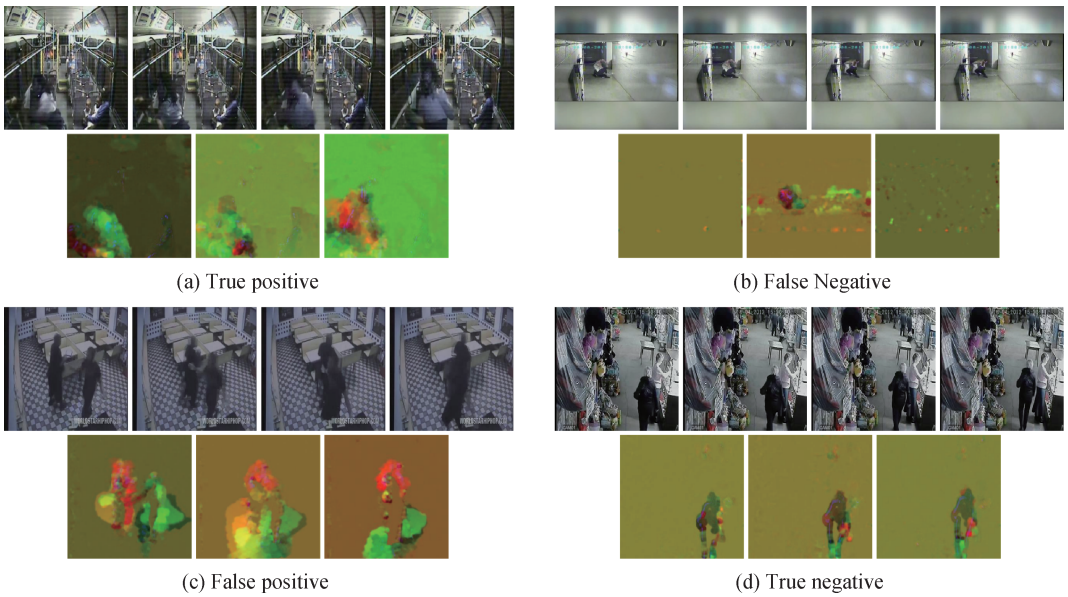
The data from the confusion matrix enables the calculation of various evaluation indicators. Table 5 presents the Accuracy, Precision, Recall, and F1-score of the proposed model on the Hockey and RWF datasets. On datasets with an equal number of positive and negative samples, Accuracy and F1-score tend to be close, both

serving as comprehensive indicators for evaluating the model. Precision and Recall, on the other hand, can reflect the tendency of the model's detection performance. With the samples in the confusion matrix, we can specifically analyze the detection performance and application scenarios of the proposed model.

Table 5 Detailed evaluation results on Accuracy, Precision, Recall, and F1-score.

Method	Dataset	Accuracy	Precision	Recall	F1-score
[35]	Hockey	98.5%	0.99	0.98	0.985
	RWF2000	88.2%	0.849	0.930	0.888
[38]	RWF2000	90.4%	0.9	0.91	0.9
Ours	Hockey	97.5%	0.98	0.97	0.975
	RWF2000	91%	0.923	89.5	0.909

We selected representative samples from the RWF2000 dataset, as shown in Figure 6. The characters in the True Positive samples appear blurred and incomplete, while the True Negative samples contain complex backgrounds and irrelevant characters. Our proposed model can accurately evaluate such samples, demonstrating high robustness to complex scenes. In such video samples, the splicing input can extract key motion information and remove background noise. However, upon observing the test samples with incorrect predictions, we found that nearly half of the False Negative samples, despite being 5 seconds in duration, consist of only a few frames of images, where the images remain unchanged over an extended period. Consequently, the splicing input contains numerous blank frames, posing a significant challenge to models that rely on motion information to identify violent behavior. In False Positive samples, there is an instance where two people intersect quickly. Due to the inherent limitations of spatial depth perception in 2D images, this instance is prone to being misidentified as violent behavior. These observations shed light on the strengths and limitations of our proposed model when dealing with specific types of samples. From this perspective, our proposed model exhibits a higher sensitivity to violent behavior. This sensitivity proves beneficial in practical applications for violence detection, as it reduces the likelihood of overlooking actual violent behavior and serves as an alert for potential dangers such as rapid intersections.



(a) True positive (b) False Negative

(c) False positive (d) True negative

Fig.6 Video frames and splicing inputs of confusion matrix samples

5 Conclusions

Our proposed violence detection method integrates deep and shallow features through a fusion approach, where the optical flow and grayscale frame difference are extracted from the video as inputs to the network. Spatial features are obtained using the pre-trained MobileNet V3, while the fusion of deep and shallow features enhances the abstract representation of motion features in the deeper layers of the network. Through the implementation of the channel attention mechanism to redistribute the weight of the deeper features and the integration of ConvLSTM to capture temporal dependencies, the temporal correlation among video sequences is enhanced. Ablation studies validate the effectiveness of the proposed method. The splicing input is beneficial in enabling the network to learn motion information directly from the video, enhancing its suitability for violence detection compared to raw images. Moreover, the deep and shallow feature fusion module significantly improves the model's performance by optimizing pre-trained MobileNet and effectively

integrating different levels of features. The channel attention module and ConvLSTM adaptively assign feature weights in channel and temporal dimensions, enhancing the model's ability to capture relevant information. An analysis of the model's prediction examples underscores the promising performance of our proposed method in real-world scenarios, affirming its suitability and practicality for violence detection applications.

Author Contribution:

Lin'en Liu: Methodology, Software, Validation, Resources, Datacuration, Writing. Xuguang Zhang: Conceptualization, Methodology, formal analysis, Supervision, Validation, Resources, Writing.

Funding Information:

This research received no external funding

Data Availability:

The datasets analyzed during the current study are available in the publicly archived datasets:

Hockey Fight: http://visilab. etsii. uclm. es/personas/oscar/FightDetection/index.html

Movies: http://visilab. etsii. uclm. es/personas/oscar/FightDetection/index.html

RWF2000: https://github. com/mchengny/RWF2000-Video-Database-for-Violence-Detection

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 05 September 2024; Accepted 05 December 2024; Published online 31 December 2024

References

[1] S. Zhang, X. Qin, F. Zhen, Y. Huang, and Y. Kong, "Do surveillance cameras improve perceived neighborhood safety? *a case study of nanjing* , china," Cities, vol. 140, p. 104423, **2023**.

[2] J. Laufs, H. Borrion, and B. Bradford, "Security and the smart city: A systematic review, " Sustainable Cities and Society, vol. 55, p. 102023, **2020**.

[3] Zhang, W., He, P., Wang, S. et al. A Dynamic Convolutional Generative Adversarial Network for Video Anomaly Detection. *Arab J Sci Eng* 48, 2075-2085 (**2023**).

[4] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," *IEEE Access* , vol. 7, pp. 107 560-107 575, **2019**.

[5] Tran D., Wang H., Torresani L., Ray J., LeCun Y., & Paluri M. (**2018**). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6450-6459).

[6] Lipton Z C, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning[J]. *arXiv preprint* arXiv:1506.00019, **2015**.

[7] 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* 46, 8549 – 8563 (**2021**).

[8] He, Kaiming, et al. "Deep residual learning for image recognition. " *Proceedings of the IEEE conference on computer vision and pattern recognition* . **2016**.

[9] Kumar M., Biswas M. Abnormal human activity detection by convolutional recurrent neural network using fuzzy logic. *Multimed Tools Appl* (**2023**).

[10] Ullah, W., Ullah, A., Haq, I. U. et al. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimed Tools Appl* 80, 16979-16995 (**2021**).

[11] S. T. Sarcar and M. A. Yousuf, "Detecting Violent Arm Movements Using CNN-LSTM, " *2021 5th International Conference on Electrical Information and Communication Technology (EICT)* , Khulna, Bangladesh, **2021**, pp. 1-6.

[12] I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa and H. S. Nawaz, "A Frame-Based Feature Model for Violence Detection from Surveillance Cameras Using ConvLSTM Network, " *2021 6th International Conference on Image* , Vision and Computing (ICIVC), **2021**, pp. 55-60.

[13] S. K. Parui, S. K. Biswas, S. Das, M. Chakraborty and B. Purkayastha, "An Efficient Violence Detection System from Video Clips using ConvLSTM and Keyframe Extraction, " 2023 11th International Conference on Internet of Everything, EngineeringMicrowave, Communication and Networks (IEMECON), Jaipur, India, **2023**, pp. 1-5.

[14] Qu, W., Zhu, T., Liu, J. et al. Correction to: A time sequence location method of long video violence based on improved C3D network. *J Supercomput* 79, 1158 (**2023**).

[15] Nasaruddin, N., Muchtar, K., Afdhal, A. et al. Deep anomaly detection through visual attention in surveillance videos. *J Big Data* 7, 87 (**2020**).

[16] Carreira J., & Zisserman A. (**2017**). *Quo vadis, action recognition? a new model and the kinetics dataset* . In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[17] Ehsan, T. Z., Nahvi, M. & Mohtavipour, S. M. Learning deep latent space for unsupervised violence detection. *Multimed Tools Appl* 82, 12493 – 12512 (**2023**).

[18] Freire-Obregón, D., Barra, P., Castrillón-Santana, M. et al. Inflated 3D ConvNet context analysis for violence detection. *Machine Vision and Applications* 33, 15 (**2022**).

[19] A. R. Abdali and A. A. Aggar, "*DEVTrV* **2**: Enhanced Data-Efficient Video Transformer For Violence Detection, " 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, **2022**, pp. 69-74.

[20] Farnebäck G. (**2003**). Two-Frame Motion Estimation Based on Polynomial Expansion. In: *Bigun, J* ., Gustavsson, T. (eds) Image Analysis. SCIA **2003**. Lecture Notes in Computer

Science, vol 2749. Springer, , HeidelbergBerlin.

[21] Nemade, Neeta Anil and Vinaya V. Gohokar. "*Comparative Performance Analysis of Optical Flow Algorithms for Anomaly Detection." SSRN Electronic Journal* (**2019**): n. pag.

[22] Howard, Andrew, et al. "Searching for mobilenetv3. " *Proceedings of the IEEE/CVF international conference on computer vision* . **2019**.

[23] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. *Advances in neural information processing systems* , **2015**, 28.

[24] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, **2020**, pp. 11531-11539.

[25] E. B. Nievas, O. D. Suarez, G. B. Garc´ıa, and R. Sukthankar, "Violence detection invideo using computer vision techniques." International conference on Computer analysis of images and patterns.Springer, **2011**.

[26] C. Ming, K. Cai, and M. Li, "RWF-2000: an open large scale video database for violence detection. " 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, **2021**.

[27] T. Hassner, Y. Itcher and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, **2012**, pp. 1-6.

[28] S. Das, A. Sarker and T. Mahmud, "Violence Detection from Videos using HOG Features, " *2019 4th International Conference on Electrical Information and Communication Technology (EICT)* , Khulna, Bangladesh, **2019**, pp. 1-5.

[29] A. Jain and D. K. Vishwakarma, "Deep NeuralNet For Violence Detection Using Motion Features From Dynamic Images, " **2020** Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, **2020**, pp. 826-831.

[30] N. Honarjoo, A. Abdari and A. Mansouri, "Violence detection using pre-trained models," *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)* , Kashan, Iran, **2021**, pp. 1-4.

[31] D. G. C. Roman and G. C. Chávez, "Violence Detection and Localization in Surveillance Video," **2020** 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Porto de Galinhas, Brazil, **2020**, pp. 248-255.

[32] S. Jianjie and Z. Weijun, "Violence Detection Based on Three-Dimensional Convolutional Neural Network with Inception-ResNet, " **2020** IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, **2020**, pp. 145-150.

[33] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir and M. Farazi, "Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM, " *2021 International Joint Conference on Neural Networks (IJCNN)* , Shenzhen, China, **2021**, pp. 1-8.

[34] C. Pan and S. Fei, "Violence detection based on attention mechanism," *2022 41st Chinese Control Conference (CCC)* , Hefei, China, **2022**, pp. 6438-6443.

[35] F. U. M. Ullah et al., "AI-Assisted Edge Vision for Violence Detection in IoT-Based Industrial Surveillance Networks," *in IEEE Transactions on Industrial Informatics* , vol. 18, no. 8, pp. 5359-5370, Aug. 2022.

[36] L. Zhou, "End-to-End Video Violence Detection with Transformer," **2022** 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, **2022**, pp. 880-884.

[37] N. D ¨ undar, A. S. Kec ¸eli, A. Kaya, and H. Sever, "A shallow*3d convolutional neural network for violence detection in videos* ," Egyptian Informatics Journal, vol. 26, p. 100455, **2024**.

[38] H. Mohammadi and E. Nazerfard, "Video violence recognition and localization using a semi-supervised hard attention model," *Expert Systems with Applications* , vol. 212, p. 118791, **2023**.

[39] G. Garcia-Cobo and J. C. SanMiguel, "Human skeletons and change detection for efficient violence detection in surveillance videos, " *Computer Vision and Image Understanding* , vol. 233, p. 103739, **2023**.

[40] T. Xiang, H. Pan, and Z. Nan, "Video violence rating: A large-scale public database and a multimodal rating model," IEEE Transactions on Multimedia, pp. 1-12, **2024**.

[41] Honarjoo N., Abdari A. & Mansouri A. Violence detection in compressed video. *Multimed Tools Appl* (**2024**).

[42] Veltmeijer E., Franken M. & Gerritsen C. Real-time violence detection and localization through subgroup analysis. *Multimed Tools Appl* (**2024**).