A Baseline for Violence Behavior Detection in Complex Surveillance Scenarios

Yingying Long
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: longyingying@st.xatu.edu.cn

Zongxin Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: zongxinwang@yeah.net

Hanzhu Wei
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: weihanzhu@st.xatu.edu.cn

Xiaojun Bai
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: baixiaojun@st.xatu.edu.cn

***Abstract*—Violence detection can improve the ability to deal with emergencies, but there is still no data set specifically for violence detection. In this work, we propose VioData, a datasets specialized for detection in complex surveillance scenarios, and to more accurately assess the efficacy of these datasets, we propose a violence detection model based on target detection and 3D convolution. The model consists of two key modules: spatio-temporal feature extraction module and spatio-temporal feature fusion module. Among them, the spatio-temporal feature extraction module consists of a spatial feature module that extracts key frames using ordinary convolutional networks and a temporal feature extraction module that establishes temporal features using 3D convolution. The spatio-temporal feature fusion module Channel Fusion and Attention Mechanism (CFAM) fuses the temporal and spatial features. The experimental results indicate that the precision of the suggested detection model on UCF101-24, JHMDB behavioral detection datasets, and our proposed violence detection datasets, VioData, is improved compared to other violence detection models, which not only verifies the validity of the datasets, but also provides a baseline for the subsequent research and improvement in this area.**

Keywords-Violent Behavior Detection; Datasets; Spatio-temporal Feature; Target Detection; Feature Fusion

I. INTRODUCTION

Violent behavior is defined as the use of force and other means to harm oneself or others, and violent behavior detection can serve as one of the roles to meet the growing public safety needs. Utilizing deep learning technologies in the domain of violent behavior detection can capture eligible violent behaviors from cameras and alert the police, which is a useful tool for public security officers' daily tasks.

However, violent behaviors mostly occur outdoors, and in complex surveillance scenes with large field of view outdoors, the small size of the human target makes it challenging to locate the important parts of the body, many occlusions, and the complex background, which poses a great challenge to the detection of violent behaviors. In the existing public behavior detection datasets UCF101-24 and JHMDB, which contain 45 categories of more common behaviors, there is no violence detection datasets specifically for complex surveillance scenes. Moreover, most of the existing behavior detection algorithms use a two-stage strategy, such as SlowFast [9] and other candidate areas are initially generated by two-stage detection algorithms, and then finally perform feature extraction and classification on the

candidate regions to ultimately determine the behavioral categories and locations. However, two-stage algorithms have been difficult to apply in complex surveillance scenarios, firstly, the method of obtaining candidate frame sequences through the detection algorithm cuts off the potential relationship between people and people, people and background, etc. Finally, the operation of analyzing all detected people is challenging to fulfill the real-time requirements in reality.

Therefore, this paper collects publicly available surveillance videos of public places and takes them as the research object, and uses them as the raw data to produce a set of violence detection datasets, VioData, which is specialized in complex surveillance scenes; and offers a violence detection module utilizing target identification and three-dimensional convolutional networks. and target detection for accomplishing the violence detection task more efficiently. The module integrates the spatio-temporal feature data of the video sequence and extracts the spatial properties of the key frames through ordinary convolutional network, extracts temporal characteristics from the video using a 3D convolutional network, and finally fuses the spatio-temporal features through spatio-temporal feature fusion network. In addition, the module UCF101-24, the JHMDB datasets, and the VioData datasets constructed in this paper on which extensive experiments were carried out, and the experimental findings verify the effectiveness of the datasets and the module's ability to produce competitive outcomes in the detection of violent behavior in complex outdoor scenes. The main contributions of this paper are as follows:

- VioData, a datasets specialized for violence detection.
- Because of the occlusion phenomenon in complex violent behavior scenes, a temporal feature extraction network is proposed in this paper. which introduces 3D Convolutional Block Attention model (3D-CBAM) attention mechanism and spatio-temporal depth separable convolution to better utilize the information between consecutive frames to better extract the features in the video sequences, and to improve how the network perceives the foreground features; secondly, to detect the aggressive behavior more precisely, the introduced Atrous Spatial Pyramid Pooling (ASPP) model is introduced in order to more accurately detect violent acts, and the fusion of feature maps of different sensory fields is obtained by utilizing different scales of convolution.
- In order to naturally fuse spatio-temporal information for a later, more precise identification of aggressive behavior, a spatio-temporal feature fusion module was designed.

II. Related Work Research

We will review the work related to behavioral detection datasets and review the work on techniques used for behavioral detection from four perspectives: behavioral detection based on traditional features, behavioral detection based on recurrent neural networks, behavioral detection based on multi-stream neural networks, and behavioral detection based on three-dimensional convolutional networks.

A. Behavioral detection datasets

Behavior detection datasets typically contain data collected from sources such as videos, sensors, etc. And are used to train and test algorithms for recognizing and analyzing human behavior. The UCF101 [1] datasets is among the biggest datasets of human behavior that are currently accessible, containing 101 action categories, almost 13,000 video snippets, totaling 27 hours of footage. Real user-uploaded films with crowded backdrops and camera motions make up the database. HMDB with Joint Annotation (JHMDB) [2] datasets A subset of the Human Metabolome Database (HMDB) [3] datasets contains 21 action categories, each involving the movement of a single character. The dataset was annotated with 2D joint model, providing information on the character's pose, optical flow, and segmentation for analyzing action recognition algorithms. The Kinetics [4] datasets is a human action video datasets introduced by DeepMind that contains 400 human action categories, each with 400 video snippets, each lasting roughly 10 seconds, from different YouTube videos. The dataset covers a wide range

of action categories, including human-object interaction and human-human interaction.

B. Behavior detection based on traditional features

Before the popularization of deep learning techniques, researchers used traditional features to process image information. The technique mostly included manually removing characteristics from video frames, which were then fed into support vector machines and decision trees for further behavioral analysis and identification. Xu [5] et al. suggested a technique for detecting violent videos that uses sparse coding and MoSIFT characteristics. Initially, the low-level description of the video is extracted using the MoSIFT algorithm, then feature selection is performed by Kernel Density Estimation (KDE) to eliminate noise, and finally the selected MoSIFT features are further processed using a sparse coding scheme to obtain highly discriminative video features. Febin [6] proposed a new descriptor Motion Boundary SIFT (MoBSIFT) to more effectively identify the characteristics of violent actions in the video. This module is able to filter out the random motions in the nonviolent behaviors, and represent and classify the violent videos by sparse coding technique, which has high accuracy and robustness in detecting violent behaviors.

C. Recurrent neural network based behavior detection

By receiving the hidden state of the preceding moment, a recurrent neural network (RNN) may model the frames in a movie as an ordered sequence, which affects the state of the next moment, and the extracted temporal features are able to express human behavior. With networks like Long Short-Term Memory (LSTM), this behavior detection technique first extracts spatial data from the ordered sequence of frames, and then it goes on to extract temporal features from the video. Sudhakaran [7] proposed ConvLSTM, which aggregates frame-level violent behavioral features in the video by capturing the spatio-temporal features and captures the differences between consecutive frames by computing the motion changes, which reduces the amount of data to be processed. Liang [8] et al. used GhostNet and ConvLSTM to construct a long-term recurrent convolutional network and introduced a multiple attention mechanism in the video preprocessing stage to enhance the attention to the key information in the video, which improves the ability of detecting violent behavior in the video.

D. Behavior detection based on multi-stream neural networks

Multi-stream neural networks usually have many branches, before employing a classifier to identify behaviors, each branch independently extracts many feature streams from a large number of samples and aggregates the extracted features. Feichtenhofer [9] et al. designed a SlowFast network based on frame rate speed. The network contains two paths, Slow path and Fast path, to extract spatial semantic information and motion information at lower and higher frame rates, respectively, to enhance behavior detection. Next, Okan [10] proposed a multi-modal parallel module You Only Watch Once (YOWO) based on a dual channel structure. The network has two branches: one uses 2D-CNN to extract the spatial properties of key frames, while the other uses 3D-CNN to extract the spatio-temporal features of the video segment made up of earlier frames, and finally, fuses the features using channel fusion and the attention mechanism to perform frame-level detection for behavioral Localization of actions. Li [11] et al. suggested a novel technique for detecting violence based on a multi-stream detection model, which combines three distinct streams—a temporal stream, a local spatial stream, and an attention-based spatial RGB stream—to improve the performance of violent behavior recognition in videos. Islam [12] et al. suggested an effective dual-stream deep learning architecture using pre-trained MobileNet and LSTM (SepConvLSTM), in which one stream manages frame background suppression and the other handles frame differences between neighbors. In order to provide discriminative features that aid in differentiating between violent and nonviolent activities, a straightforward input preprocessing technique highlights moving objects in the frames while suppressing the nonmoving background and recording the inter-frame actions.

E. Behavior detection based on 3D convolutional networks

Conventional 2D convolutional neural networks that have been trained on single-frame images are unable to reflect the correlation between consecutive frames, while 3D convolutional networks are able to directly extract frames from the video, and then fed into 3D CNNs to extract the spatio-temporal features in the frame sequences, the network learns the characterization of the behaviors in the video after multilayered convolutional and pooling operations, and accurately detects the behaviors in the video, and it is currently an important research direction. Carreira [13] based on the Inception network and extended it from 2D to 3D, proposed the network Inflated 3D ConvNet (I3D) which is able to process video data for behavioral detection. Direkoglu [14] computes optical flow vectors for every frame to produce a motion quantum image (MII), It is then used to train a Convolutional Neural Network (CNN) to identify abnormal behavioral events in a crowd. The proposed MII is mainly based on the optical flow magnitude and angular difference calculated from the optical flow vectors in consecutive frames, which helps to distinguish between normal and abnormal behavior. Dong [15] et al. suggested the attentional residual 3D network (AR3D) and the residual 3D network (R3D), which were model ed by upgrading the current 3D CNNs by adding the residual structure and attention mechanism, The behavior detection performance of the model has been improved in different degrees. Li [16] et al. establish a 3D-DenseNet dense connectivity model , extract spatio-temporal features using 3D-DenseNet algorithm, redistribute the weights of each feature using the Squeeze-and-Excitation Networks (SENet) channel attention model , and then use the transition layer sampling, and then pass the outcomes to the fully connected layer using the global average pooling technique to finish the violence detection task. XU [17] et al. proposed the SR3D algorithm, which adds a BN layer before the 3D convolutional operation and presents the ReLu activation mechanism to enhance the network's learning capabilities while, extends the SE attention mechanism to 3D by introducing it into the 3D convolutional model and boosts the weights of the important channels, which improves the ability to detect the human behaviors in the video in the network.

III. VIOLENCE TEST DATASETS PRODUCTION

Because there are no samples of datasets dedicated to violence detection in the current public datasets in the field of video behavior detection such as UCF101-24, JHMDB and Kinectics. Therefore, in this paper, we produce VioData, a violence detection datasets specialized for complex surveillance scenarios.

First, this paper collects about 1500 video clips of violent behavior from publicly available real surveillance video data.

Second, since the length of the collected surveillance videos varies between 1-10 minutes and there are not many clips in which violent behaviors occur, the collected surveillance videos are manually cropped to segment the videos into short videos of violent behaviors of about 10 seconds. Then, the obtained short videos were subjected to frame extraction, before extracting the frames, the videos were converted into easily labeled RGB image sequences and the blurred images were discarded, and the extracted video frames were deposited into a separate folder to obtain a separate clip of violent behavior using a frequency of 1 frame every 5 frames.

Finally, the human targets of violent acts in the video are labeled with frame-level truth frames using the LabelImg tool, based on the collected violent actclip clips, the manual labeling method is used, the violent act targets are labeled with rectangular frames, and the targets with more than 50% occlusion are not labeled, and the violent act targets of the part-frame Pictures are labeled as Fig.1 illustrates. The labeled information is saved as an XML file, and the xml file contains the image file address, the truth frame coordinate information and the behavioral category of the target.

(a)abuse

(b)attack

(c)fight

Figure 1. Illustration of a sample of labeled acts of violence

IV. Methodology Of This Paper

The framework of the violence detection module is shown in Fig.2, the framework has two branches of inputs, the output is a series of video frames with a violence detection frame containing the outcomes of the violence category, while the first branch consists of a series of video frames and the second branch consists of extracted keyframes. There are three modules in the module: one for spatiotemporal feature extraction, one for spatiotemporal feature fusion, and the structure of the spatio-temporal feature extraction model consists of an I3D network and a CSPDarkNet-Tiny network for extracting spatial features. The 3D convolution-based I3D network video is used for temporal modeling and for extracting temporal features; the CSPDarkNet-Tiny network model is the 2D features of the keyframes and is used for extracting the spatial features of the keyframes. The temporal and spatial feature fusion model integrates the feature information of the two branches and filters the valid information among them, lastly, to obtain the violence detection findings, the fused feature map results are input into the prediction head output.

A. Spatio-temporal feature extraction module

1) Timing feature extraction module

Violence detection for complex surveillance scenarios requires high real-time modeling, and occlusion phenomena are likely to occur in the violence scenarios. The 3D Inception (3D) Inc model in the Inflated 3D ConvNet (I3D) network uses ordinary 3D convolution, but its computational overload makes it difficult to perform real-time violence detection. The original Inflated 3D ConvNet (I3D) network is prone to omission and false detection when detecting violence with occlusion phenomenon. Therefore, in this work, according to the features of the original I3D network structure, spatio-temporal depth-separable convolution and 3D-CBAM attention are introduced to improve both efficiency and accuracy.

In terms of real-time, after frame-by-frame convolution operation, the spatio-temporal information is combined by point-by-point convolution to extract higher-level feature representation in real time. The improved 3D Inc reduces the computational effort of the 3D Inc module exponentially by replacing the standard $3 \times 3 \times 3$ convolution in the middle two branches with spatio-temporal depth-separable convolutions of $1 \times 3 \times 3$ and $3 \times 1 \times 1$ shapes. The 3D Inc module finally fuses the features of the four branches. The structural diagram of the optimized 3D Inc network is shown in (c) in Fig.2.

In terms of accuracy, the Convolutional Block Attention model (CBAM)[18], which aggregates the temporal dimension information based on CBAM, is introduced in this study since the temporal information in the video sequences cannot be properly utilized. The structure diagram of 3D-CBAM is shown in (b) in Fig.2. The Channel Attention model (CAM) and the Spatial Attention model (SAM) make up 3D-CBAM. The Channel Attention model processes the input feature map F3D to produce the channel weight vector. which is multiplied with F3D to obtain the

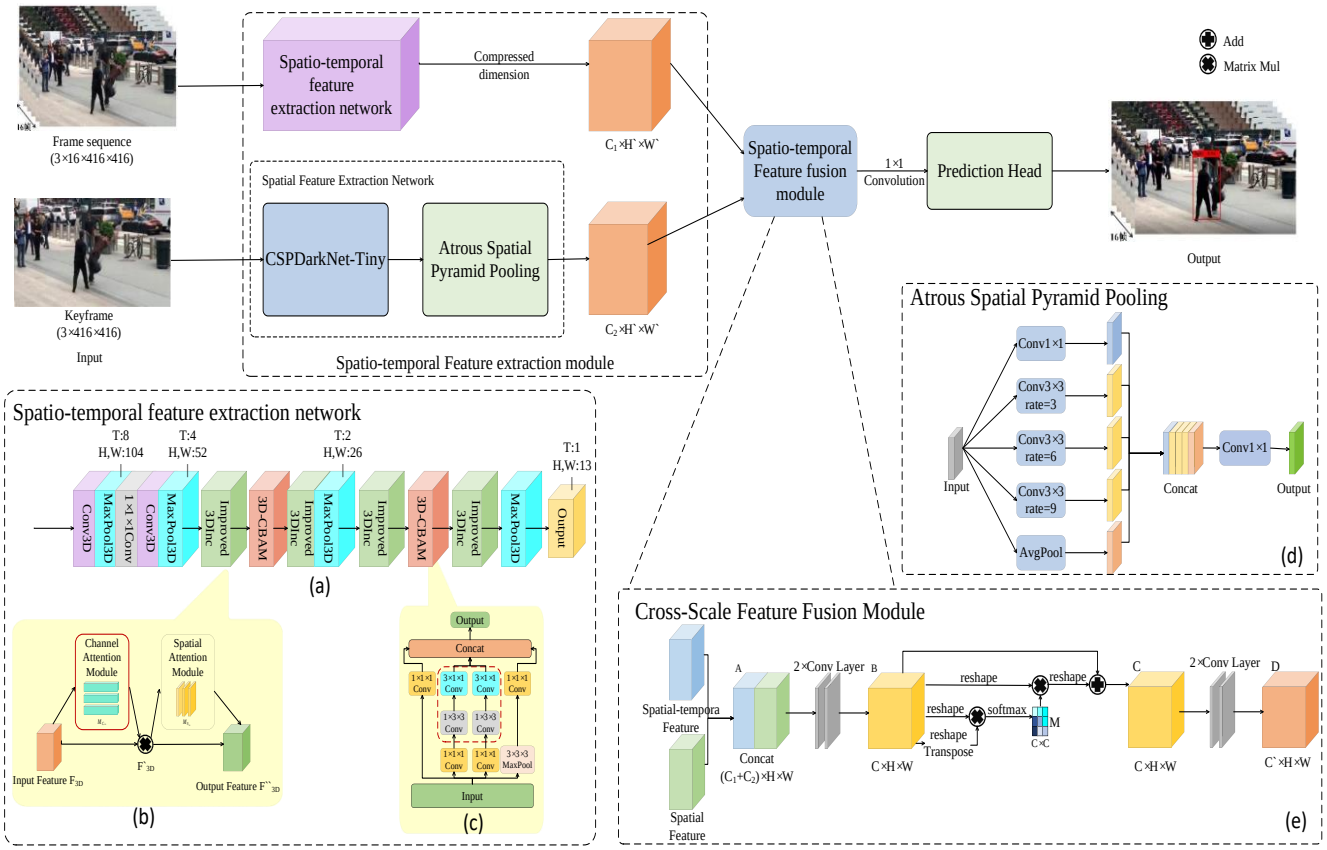

Figure 2. The violence detection algorithm's framework is displayed in Fig.2. The model for extracting spatio-temporal features and the spatio-temporal feature fusion module make up the majority of the framework. The spatio-temporal feature extraction model is composed of the temporal feature extraction model and the spatial feature extraction module, and the I3D network is the network structure of the temporal feature extraction model, as illustrated in (a); (b)(c) are the 3D-CBAM Attention Mechanism and 3D Inception (3D Inc) module, respectively. The Atrous Spatial Pyramid Pooling (ASPP) model is added at the end of the spatial feature extraction model, which has the CSPDarkNet-Tiny network as its network structure, which is shown in (d), where rate denotes the expansion rate of the null convolution. atrous Spatial Pyramid Pooling (ASPP) has five branches, including one ordinary convolutional branch, three null convolutional branches, and one global average pooling branch; (e) shows the overall structure of Channel Fusion and Attention Mechanism(CFAM); D is the final output feature map of CFAM, and C1 and C2 are the number of feature map output channels for the I3D network and the ASP module, respectively.

F'_{3D} weighted feature map. The Spatial Attention model then processes F'_{3D} to get the spatial weight, which is then multiplied by the feature F'_{3D} to get the final feature F''_{3D}, which combines spatial and channel attention. channel and spatial focus of the F''_{3D} feature. The following is the computational expression for 3D-CBAM:

$$F'_{3D} = M_{C_{3D}}\left(F_{3D}\right) \otimes F_{3D} \tag{1}$$

$$F''_{3D} = M_{S_{3D}}\left(F'_{3D}\right) \otimes F'_{3D} \tag{2}$$

where $M_{S_{3D}}$ represents the spatial attention, and $M_{C_{3D}} \in R^{C \times D \times 1 \times 1}$ represents the channel attention. D is the number of frames in the video sequence frame, while C is the number of feature map channels. In Fig.2, the enhanced I3D network structure is displayed in (a).

2) Spatial feature extraction module

Wang [19] et al proposed CSPDarkNet combines the features of Cross Stage Partial Network (CSP) structure and DarkNet framework, which is able to maintain or even improve the capability of CNN while reducing the amount of computation. In this paper, considering the scenario of violence detection, we need to use an efficient and lightweight network, so we chose a lightweight CSPDarkNet network, CSPDarkNet-Tiny, its network structure is shown in Fig.3, but because of the violence detection method, the lightweight network may lead to insufficient computational power to deal with occlusion or

background complexity, which leads to the decrease of accuracy. Therefore, this paper introduces CSPDarkNet-Tiny's last layer is supplemented with the Atrous Spatial Pyramid Pooling (ASPP) module. The ASPP input feature maps are branched through five null convolutions to obtain feature maps with five different sensory fields, which are spliced and fused along the channel dimensions, and then adjusted using a 1× 1 convolutional number of channels to acquire more specific visual information. Fig.2(d) displays the ASPP model's structure.

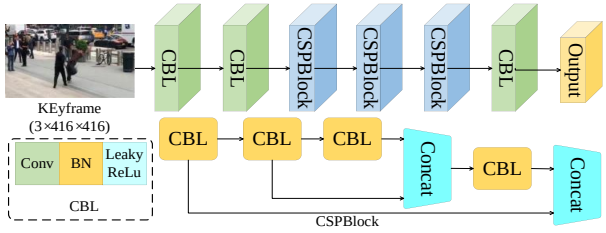

Figure 3. CSPDarkNet-Tiny Network Overall Structure

B. Spatio-temporal feature fusion module

The temporal fusion attention model(TFAM) [19] is an attention mechanism module for improved video object detection, which improves object representation by combining multi-frame and single-frame attention modules and dual-frame fusion modules, but it has too much computation and weak generalization ability, which is not conducive to the application of violent behavior detection. Therefore, Channel Fusion and Attention Mechanism (CFAM) is introduced in this paper to effectively integrate the temporal features obtained by I3D network with the spatial features obtained by CSPDarkNet-Tiny network to record the inter-channel dependencies. Fig.2(e) displays the CFAM model's structure diagram.

The following is how feature fusion specifically works: firstly, the feature maps obtained from the first two networks are spliced to obtain the feature map $A \in R^{(C1+C2)\times H\times W}$, then the correlation between the feature maps is captured using the local receptive fields of the convolutional layers, and the correlation feature map $B \in R^{C\times H\times W}$ is produced by passing the feature map A through two convolutional layers. Since direct correlation calculation would make the computation complicated, a reshaping operation is performed on B to obtain a reshaped feature map F. The elements of each channel in the feature map are converted into one-dimensional vectors to simplify the computation. The expression is as follows:

$$B \in R^{C\times H\times W} \xrightarrow{\text{vectorization}} F \in R^{C\times H\times W} \tag{3}$$

First, the resulting feature map F is dot-producted with its transposed feature map F^T to obtain a covariance matrix $G \in R^{C\times N}$, where $N=H\times W$. This matrix reveals the correlation between different features. Its expression is as follows:

$$G = F \times F^T \tag{4}$$

$$G_{i,j} = \sum_{k=1}^{N} F_{ik} \times F_{jk} \tag{5}$$

Where $G_{i,j}$ represents the inner product between the feature map F and F^T. After that, the resulting matrix G is subjected to *softmax* operation to generate the channel attention feature map $M \in R^{C\times C}$. The *softmax* function is able to transform the values between the range of 0-1, which represents the attention weight of each position. the expression of M feature map is as follows:

$$M_{i,j} = \frac{e^{G_{ij}}}{\sum_{j=1}^{C} e^{G_{ij}}} \tag{6}$$

In order for the attention map M to have an effect on the original feature map, the matrix F' is obtained by dot-product multiplication of M with the reshaping matrix F, which makes the features of the parts with high weights more prominent. Then F' is reshaped to $F'' \in R^{C\times H\times W}$ of the same size as B.

$$F' = M \times F \tag{7}$$

To alleviate the gradient vanishing problem and accelerate the model convergence, F'' is multiplied with the hyperparameter α and superimposed with the feature map B using the expression in (8) to get the feature map $C \in R^{C\times H\times W}$, the final spatio-temporal feature map $D \in R^{C\times H\times W}$ with the

attention weights is obtained by consecutively applying two convolutions to the resultant feature map *C*.

$$C = \alpha \times F'' + B \tag{8}$$

C. Loss function

The loss function proposed in this paper contains three components: classification prediction loss L_{cls}, localization loss L_{rect}, and confidence loss L_{obj}.

The classification prediction loss formula is as follows:

$$\mathrm{y_i} = \mathrm{sigmoid}\left(\mathrm{x_i}\right) = \frac{\mathrm{e}^{\mathrm{x_i}}}{\sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{e}^{\mathrm{x_n}}} \tag{9}$$

$$L_{cls}\left(y, y_i\right) = -\frac{1}{N}\sum_{n=1}^{N} L_{BCE_{cls}} \tag{10}$$

$$L_{BCE_{cls}} = y \times \log\left(y_i\right) + (1-y) \times \log\left(1-y_i\right) \tag{11}$$

where x_{i} is the category's projected value and *N* is the total number of categories in the datasets, y_i is the current category probability, and *y* is the true value of the current category.

The localization loss formula is as follows:

$$v = \frac{4}{\pi}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{12}$$

$$\alpha = \frac{v}{1 - \mathrm{IoU}\left(B, B_{gt}\right) + v} \tag{13}$$

$$L_{rect}\left(B, B_{gt}\right) = CIoU\left(B, B_{gt}\right) - \frac{p^2(B, Bgt)}{c^2} - \alpha \tag{14}$$

Where w^{gt} as well as h^{gt} are the target real frame's width and height, the target predicted frame's width and height, *v*, which is the value of the projected frame normalized by extrapolating the width-to-height ratio of the predicted frame to the actual frame, and p^2, which is the distance between the predicted frame's centroid and the real frame's centroid, where α represents the balance between the loss resulting from the measurement of aspect ratio and the loss due to IoU. The confidence loss is publicized as follows:

$$L_{obj}\left(C, C_i\right) = -\frac{1}{N}\sum_{n=1}^{N} L_{BCE_{obj}} \tag{15}$$

$$L_{BCE_{obj}} = C \times \log\left(C_i\right) + (1-C) \times \log\left(1-C_i\right) \tag{16}$$

where *C* is the current grid region's confidence, C_i is the expected value of confidence, and *N* is the number of feature points.

The three loss functions above are integrated to get the total loss function with the following formula:

$$L = \alpha_1 \times L_{cls} + \alpha_2 \times L_{rect} + \alpha_3 \times L_{obj} \tag{17}$$

To ensure that the weights of the various loss terms are balanced, the hyperparameters α_1, α_2 and α_3 are set. where α_1 has a value of 0.4, α_2 has a value of 0.3 and α_3 has a value of 0.3.

V. Experiments And Analysis Of Results

A. Experimental setup

The Kinectics datasets are used to train the model suggested in this research, and the custom datasets VioData are used to refine it.

In order to be able to enrich the training set and make the model better acquire the effective features in the video frames, three data enhancement operations are adopted in this paper, including horizontal flipping, random scaling, and color enhancement. The data enhancement operations expand the datasets, reduce overfitting, enhance the generalization ability of the model, and improve the robustness of the model.

The training settings are displayed in Table Ⅰ below.

TABLE I. Parameter settings in network training

Parameter	Setting
Initial Learning Rate	0.001
Epoch	230

Parameter	Setting
ReSize	（416,416）
Weight Decay	0.0005
Optimizer	Adam

Configure the model parameters to be saved once per ten iterations while the model is being trained, and save the output of the training loss and validation loss once at the completion of each epoch. When the loss iteration reaches 180 rounds, the training loss is still decreasing, but the loss of the validation set starts to rise, indicating that the model has been overfitted, so the model parameters at the end of the 180th epoch are saved as the optimal parameters.

The effect of the network model for violence detection on the VioData datasets is visualized as shown in Fig.3, where a video of a violent act is subjected to model inference to obtain the location of the violent act and its category, proving the effectiveness of the VioData datasets.

Figure 4. Violence detection results

B. Experimental results and analysis

To compare with other violence detection techniques and show the efficacy of the suggested enhanced modules in the suggested violence detection model, we conducted numerous tests in this work. The experiments are conducted on three datasets (UCF101-24, JHMDB, and VioData).

1) Experimental result and analysis

In order to confirm the model's efficacy for violence detection, the model put out in this work is contrasted with current behavioral detection techniques in this section. The following four models are chosen for comparison studies:

a) MPS [21]: this model proposes a new fusion strategy that not only fuses the appearance and optical flow information of dual-stream networks, but also includes a solution to the problem of small camera movements.

b) P3D-CTN [22]*:* the core idea of this model is to use the so-called Pseudo-3D Convolution, which is a method that combines 2D spatial convolution with 1D temporal convolution. This method can effectively extract spatio-temporal features from videos without significantly increasing the computational complexity.

c) STEP [23]*: t*his model contains two main parts, spatial refinement and temporal expansion. Each step in spatial refinement uses the regression output of the previous step to improve the quality; temporal extension focuses on improving the accuracy of action classification through the duration of the video clip.

d) YOWO [10]*: t*his architecture contains two branches, one for extracting spatial features of key frames and the other for modeling the spatio-temporal features of video clips consisting of previous frames, and finally the features obtained from the two branches are fused through the attentional mechanism and regressed for classification.

The outcomes of this comparison experiment are displayed in the Table Ⅱ:

TABLE II. RESULTS OF VIOLENCE DETECTION ACCURACY OF DIFFERENT MODELS

Method	UCF101-24	JHMDB	VioData
	mAP		
MPS	82.4%	-	85.3
P3D-CTN	-	84.0%	84.9%
STEP	83.1%	-	86.4%
YOWO	82.5%	85.7%	88.0%
ours	89.8%	88.6%	91.8%

2) Ablation experiments

In this part, we use a series of ablation experiments to assess how various network enhancements affect the effectiveness of video behavior identification.

First, we introduce the ASPP model on the CSPDarkNet-Tiny backbone network, and next, we introduce spatio-temporal depth-separable convolution in the I3D network, and further

experiments are conducted on the same datasets. The experimental results are shown in Table Ⅲ.

TABLE III. DETECTION RESULTS WITH EMBEDDED ASPP MODEL AND INTRODUCTION OF SPATIO-TEMPORAL DEPTH SEPARABLE CONVOLUTION

Network	UCF101-24	JHMDB	VioData
	mAP		
Baseline	78.5%	75.3%	78.9%
CSPDarkNet-Tiny+ASPP	80.7%	76.6%	82.0%
CSPDarkNet-Tiny+ASPP++I3D(Improved 3D Inc)	84.8%	80.4%	86.5%

Table Ⅲ makes it clear that the ASPP paradigm was introduced in the CSPDarkNet-Tiny network has an improvement of 2.2, 1.3, and 3.1 percentage points on the three datasets, respectively, which indicates that the ASPP model is effective in improving the detection accuracy of the model . By introducing spatio-temporal depth-separable convolution to improve the I3D network, the model accuracy has an improvement of about 4 percentage points on all three datasets, indicating the effectiveness of spatio-temporal depth-separable convolution in improving the detection accuracy.

Finally, we embedded the 3D-CBAM attention model in the improved I3D network and conducted experiments at different embedding locations. Table Ⅳ displays the findings of the experiment.

TABLE IV. DETECTION RESULTS OF 3D-CBAM ATTENTION MODEL EMBEDDED AT DIFFERENT LOCATIONS

Network	Embedding position	UCF101-24	JHMDB	VioData
		mAP		
I3D	-	84.4%	80.4%	86.5%
	3D Inc_1	86.1%	83.7%	89.0%
	3D Inc_2	86.7%	83.3%	88.3%
	3D Inc_3	85.9%	84.2%	89.6%
	3D Inc_1+3D Inc_2	88.2%	87.5%	90.7%
	3D Inc_1+3D Inc_3	89.8%	88.6%	91.8%
	3D Inc_2+3D Inc_3	88.0%	88.0%	91.4%
	3D Inc_1+3D Inc_2+3D Inc_3	90.0%	88.7%	92.0%

As seen in Table 3.3, the addition of the 3D-CBAM attention model has a corresponding improvement on all three datasets, and embedding more than one will give a further improvement over embedding one. Among them, adding the attention model after the first, second and third 3D Inc modules performs the best on all three datasets, but due to the consideration of the amount of parameter computation, adding the 3D-CBAM attention after the first and third 3DInc not only gives better accuracy, but also keeps the network's computation from being overly large to satisfy the requirements of video detection.

VI. CONCLUSIONS

Aiming at the problem that there is no specific violence detection data set in complex surveillance scenarios, this paper collects 1,500 violence surveillance videos in public data sets, filters and extracts the collected videos, and manually marks each frame to obtain violence detection data set VioData. This work suggests a violence detection module based on target identification and 3D convolution to deal with opacity and ambiguous human targets while detecting violence in intricate surveillance situations. This work suggests a violence detection module based on target detection and 3D convolution for detection in complex surveillance scenarios with occlusion issues and ambiguous human targets. To enhance the capacity to extract human traits from key frames, the ASPP module is incorporated into the network architecture; the 3D Inc module is improved to minimize the amount of network parameters; and by embedding the 3D-CBAM attention mechanism, the network is able to focus more on detecting the key regions of violent behavior based on the weight of the feature map. In the experimental phase, this paper first verifies whether the ASPP model is effective, followed by a comparative analysis of the 3D Inc model before and after optimization. Prior to model training, data augmentation operations are carried out on the video data to increase the model's capacity for generalization. The experimental results demonstrate that the approach suggested in this paper can successfully improve the precision of violence detection, verify the validity of the

datasets and propose benchmarks for researchers to improve the enhancement.

Considering that the experimental data is still limited, the scenes in the video data are not rich and complex enough, and the crowd violence category is not rich enough. In the future, we will continue to collect videos and look for datasets with more complex and diverse backgrounds that contain multiple violence categories.

ACKNOWLEDGMENT

This work was supported by the College Students' Innovative Entrepreneurial Training Plan Program (No. S202310702107).

REFERENCES

[1] Soomro K, Zamir A R, Shah M. UCF101: A Datasets of 101 Human Actions Classes from Videos in The Wild [J]. Computer Science, 2012.DOI: 10.48550/arXiv.1212.0402.

[2] Jhuang H, Gall J, Zuffi S, et al. Towards understanding action recognition [C] //IEEE International Conference on Computer Vision. IEEE, 2014. DOI: 10.1109/ICCV.2013.396.

[3] Wishart D S, Djoumbou F Y, Ana M, et al. HMDB 4.0: the human metabolome database for 2018 [J]. Nucleic Acids Research, 2017(D1): D1.DOI: 10.1093/nar/gkx1089.

[4] Kay W, Carreira J, Simonyan K, et al. The Kinetics Human Action Video datasets [J]. 2017.DOI: 10.48550/arXiv.1705.06950.

[5] Xu Long, Gong Chen, Yang Jie, et al. Violent video detection based on mosift feature and sparse coding [C] //2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 2014:3538-3542.

[6] Febin I P, Jayasree K, Joy P T. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm [J]. Pattern Analysis and Applications, 2020, 23(2):611-623.

[7] Sudhakaran S, Lanz O. Learning to Detect Violent Videos using Convolutional Long Short-Term Memory[C]. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017:33-34.

[8] Liang Qicheng, Li Yong, Yang Kaikai, et al. Long-term recurrent convolutional network violent Behaviour recognition with attention mechanism [J]. MATEC Web of Conferences, 2021, 336 (1): 5013.

[9] Feichtenhofer C, Fan Haoqi, Malik J, et al. SlowFast Networks for Video Recognition [C] //Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.

[10] Okan Köpüklü, Wei Xiangyu, Rigoll G. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization [J]. arXiv preprint arXiv:1911. 06644, 2019.

[11] Li Hongchang, Wang Jing, Han Jianjun, et al. A novel multi-stream method for violent interaction detection using deep learning [J]. Measurement and Control, 2020, 53(5):796-806.

[12] Islam Z, Rukonuzzaman M, Ahmed R, et al. Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM [C] //2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[13] Carreira J, Zisserman A Quo Vadis, Action Recognition? A New Model and the Kinetics datasets [J]. IEEE, 2017. DOI: 10. 1109/CVPR. 2017. 502.

[14] Direkoglu C. Abnormal Crowd Behavior Detection Using Motion Information Images and Convolutional Neural Networks [J]. IEEE Access, 2020, PP (99): 1-1. DOI: 10. 1109/ACCESS. 2020. 2990355.

[15] Dong Min, Fang Zhenglin, Li Yongfa, et al. AR3D: Attention Residual 3D Network for Human Action Recognition [J]. Sensors, 2021, 21(5):1656-1669.

[16] Li Zhan. Research on Video Violence Detection Algorithm Based on 3D Convolutional Neural Network [D]. Anhui University of Architecture, 2022. DOI: 10.27784/d. cnki. gahjz. 2022. 000160.

[17] XU Pengfei, ZHANG Pengchao, LIU Yaheng, et al. A human behavior detection algorithm based on SR3D network [J]. Computer Knowledge and Technology, 2022, 18(01):10-11. DOI: 10.14004/j.cnki.ckt.2022. 0068.

[18] Sanghyun Woo, Jongchan Park, Joon-Young Lee,In So Kweon. CBAM: Convolutional Block Attention Module. 2018.

[19] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN [C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2020. DOI: 10. 1109/CVPRW50498. 2020. 00203.

[20] Lim B, Ark S, Loeff N, et al. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting [J]. International Journal of Forecasting, 2021(1). DOI: 10. 1016/j. ijforecast. 2021. 03. 012.

[21] Alwando E, Yie-Tarng Chen, Wen-Hsien. CNN-Based Multiple Path Searchfor Action Tube Detection in Videos [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 30 (1): 104-116.

[22] Wei Jiangchuan, Wang Hanli, Yi Yun, et al. P3D-CTN: Pseudo-3D Convolutional Tube Network for Spatio-Temporal Action Detection in Videos [C] //2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 300-304.

[23] Yang Xitong, Yang Xiaodong, Liu Mingyu, et al. STEP: Spatio-Temporal Progressive Learning for Video Action Detection [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 264-272.