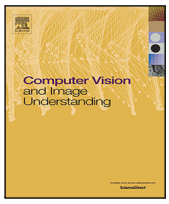
Computer Vision and Image Understanding 248 (2024) 104123

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



# A framework for detecting fighting behavior based on key points of human skeletal posture☆

Peng Zhang [a], Xinlei Zhao [b], Lijia Dong [b], Weimin Lei [a,*], Wei Zhang [a], Zhaonan Lin [a]

[a] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China
[b] Shenyang Er Yi San Electronic Technology Co., Ltd., Shenyang 110023, China

ARTICLE INFO

*Keywords:*
Violent detection
Human key points
Deep learning
Violence recognition
Automated video surveillance

ABSTRACT

Detecting fights from videos and images in public surveillance places is an important task to limit violent criminal behavior. Real-time detection of violent behavior can effectively ensure the personal safety of pedestrians and further maintain public social stability. Therefore, in this paper, we aim to detect real-time violent behavior in videos. We propose a novel neural network model framework based on human pose key points, called Real-Time Pose Net (RTPNet). Utilize the pose extractor (YOLO-Pose) to extract human skeleton features, and classify video level violent behavior based on the 2DCNN model (ACTION-Net). Utilize appearance features and inter frame correlation to accurately detect fighting behavior. We have also proposed a new image dataset called VIMD (Violence Image Dataset), which includes images of fighting behavior collected online and captured independently. After training on the dataset, the network can effectively identify skeletal features from videos and locate fighting movements. The dataset is available on GitHub (https://github.com/ChinaZhangPeng/Violence-Image-Dataset). We also conducted experiments on four datasets, including Hockey-Fight, RWF-2000, Surveillance Camera Fight, and AVD dataset. These experimental results showed that RTPNet outperformed the most advanced methods in the past, achieving an accuracy of 99.4% on the Hockey-Fight dataset, 93.3% on the RWF-2000 dataset, and 93.4% on the Surveillance Camera Fight dataset, 99.3% on the AVD dataset. And with speeds capable of reaching 33fps, state-of-the-art results are achieved with faster speeds. In addition, RTPNet can also have good detection performance in violent behavior in complex backgrounds.

## 1. Introduction

For the safety and prevention of public surveillance systems, it is necessary to detect and identify fighting behaviors in the massive video data captured through high-definition cameras. Identifying human actions in videos or images plays a crucial role in various applications, such as robots (Claudi et al., 2012; Dixit and Dhayagonde, 2014), anomaly detection (Sultani et al., 2018; Ergen and Kozat, 2019), dense scenes (Gnouma et al., 2018; Mohammadi et al., 2015), and surveillance cameras (Vosta and Yow, 2022; Garcia-Cobo and San-Miguel, 2023; Mugunga et al., 2021). At present, most of the existing combat recognition datasets are annotated for videos, such as ice hockey matches (Nievas et al., 2011), real world fights (Perez et al., 2019), movies (Chen et al., 2011; Demarty et al., 2015), and group records (Hassner et al., 2012; Robert Fisher and Crowley, 2004; Mehran et al., 2009). There is no dataset for extracting human key points to identify combat features. Therefore, to verify the importance of human key points in violence detection tasks, we collected and annotated a diverse dataset, which is called the Violent Image Dataset (VIMD).

There are still some challenges in the research of violence detection in real-life surveillance videos, such as high similarity between fighting and non-fighting actions, frequent occlusion, and complex backgrounds in group behavior. Violence detection tasks can be divided into two categories: recognition based on appearance features and recognition based on key points of the human skeleton. Utilize the appearance information in videos (Islam et al., 2021; Liu et al., 2022; Tran et al., 2018; Arnab et al., 2021) to directly use videos as input for deep neural networks. In violent video sequences, violent behavior is often concentrated in key temporal frames, so how distinguishing video keyframes is the key to solving violence detection. By introducing attention, keyframes can be highlighted, but a single mechanism for

☆ The research is supported by the 'Jie Bang Gua Shuai' Science and Technology Major Project of Liaoning Province in 2022 (No. 2022JH1/10400025), the Fundamental Research Funds for the Central Universities of China (No. N2216010).

* Corresponding author.
*E-mail address:* leiweimin@mail.neu.edu.cn (W. Lei).

https://doi.org/10.1016/j.cviu.2024.104123
Received 25 November 2023; Received in revised form 12 July 2024; Accepted 13 August 2024
Available online 21 August 2024

adding attention does not pay attention to inter-frame information. Islam et al. (2021) used optical flow information to extract features, but it also required extremely high computational costs and could not meet real-time requirements. So although such methods can effectively identify actions. However, they are not suitable for violence detection tasks in densely populated and other complex scene conditions. On the other hand, the human skeleton-based methods (Li et al., 2019a; Liu et al., 2020; Moon et al., 2021; Omarov et al., 2022; Zhang et al., 2020; Chen et al., 2021) analyze the actions and positioning of characters by extracting key points of human posture, which is more robust and performs better than the first method in identifying appearance features. However, the algorithm also has a large number of parameters and needs to address real-time issues. In this paper, we propose a novel neural network modeling framework based on features extracted from key points of human posture to achieve real-time detection of violent behavior in surveillance videos. Our main goal is to capture the motion relationship between frames through human pose key points feature extraction, capture the motion relationship between frames, and then improve the model's extraction efficiency of key frames; and maintain the model accuracy and real-time performance through the effective fusion of RGB information and skeletal information.

Our contributions are summarized as follows:

- We propose a novel neural network model framework based on human posture key points, called RTPNet. The overview of RTPNet is shown in Fig. 1; It can flexibly replace pose extractors and video comprehension level classifiers; Effectively enhancing the generalization of fighting action recognition; Unlike previous methods, the method proposed in this article takes the pose of key points as input and combines RGB information and human skeletal point information to recognize and classify fighting behavior.
- Firstly, we introduce the YOLO-Pose (Maji et al., 2022) pose extractor to extract the human skeleton sequence from the images in each frame, and introduce YOLO-Pose as a pre training method into the task of fight detection. And it was validated on the VIMD dataset. RTP-Net uses the stacked skeleton images obtained by the attitude estimator YOLO-Pose. Then, the confidence score output by YOLO-Pose was inserted into the 2DCNN network (ACTION-Net (Zhao et al., 2019)) as a weight coefficient to achieve multimodal brute force detection.
- We propose the Weight Selection Module (WSM), which uses the parameters generated from skeleton data as a weight coefficient to allocate skeletal information and RGB information. And propose a Weight Distribution Module (WDM) to further correct the error of weight coefficients, making the model pay more attention to keyframe information. Finally, the skeleton information and RGB features are connected in Keyframe Weight Assignment (KWA) to generate the final detection result.
- We collected and annotated a diverse dataset called Violent Image Dataset (VIMD), which includes online collection and surveillance video footage of fighting scenes. The scene of a fight refers to two or more people using their bodies or objects with the intention of physically harming each other. Other human interactions such as hugging, falling, and throwing objects are considered negative samples of nonviolence. The final dataset includes 2422 fighting and non-fighting images, as well as 1054 labeled skeletal information labels.
- Extensive experimental results verify that RTPNet exceeds the state-of-the-art performance. and operates in real time without sacrificing accuracy.

The rest of this paper is organized as follows. In Section 2, recent research advances related to violence detection tasks are briefly summarized. In Section 3, we present the overall architecture of the model of RTPNet and the details of each module. In Section 4, the VIMD dataset constructed in this paper and the experimental details are presented. The experimental results of RTPNet on the four datasets are analyzed in Section 5. Finally, the full paper is summarized in Section 6.

## 2. Related works

### 2.1. Appearance-based violent detection

Appearance-based violence detection Detection (Islam et al., 2021; Tran et al., 2018; Arnab et al., 2021) methods in which the appearance features of the target are extracted directly in order to achieve explicit modeling of the appearance of violent acts. Such methods have earlier used feature extraction of RGB and optical flow image information in delivery to 2D Convolutional Neural Networks (2DCNN) or 3D Convolutional Neural Networks (3DCNN (Ding et al., 2014)). Islam et al. (2021) Using optical flow information as input for 2DCNN to extract motion features. On the basis of C3D (Tran et al., 2015; Carreira and Zisserman, 2017) extended the weights of 2D convolution to 3D convolution. Zhou et al. (2017) constructed FightNet to represent complex visual violent interactions. FightNet proposes an input method for image acceleration fields. Firstly, the optical flow field is calculated using continuous frame RGB images, and the acceleration field is obtained based on the optical flow field. Sudhakaran and Lanz (2017) used convolutional long short-term memory learning to detect violent videos. In addition to the double stream 2DCNN method mentioned above, Mohtavipour et al. (2022) proposed a multi-stream CNN framework based on handcrafted extracting specific features. This method divides the extracted features into three types: spatial, temporal, and spatiotemporal streams as inputs. The spatial stream trains the network with each frame in the video to learn environmental patterns. The temporal stream consists of three consecutive frames to learn the motion patterns of intense behavior with modified optical flow differential amplitudes. In addition. A differential motion energy image with novel discriminative features was designed in spatiotemporal stream . Finally, violent behavior detection is achieved by fusing multi-stream features. These appearance-based methods (Zolfaghari et al., 2018; Li et al., 2020) can be used to recognize detailed motion features. However, it performs poorly in dealing with complex scenes and group behavior, and the model is unable to cope with different lighting background conditions.

On the other hand, some studies have focused on reducing the computational cost and parameter count of 3DCNN. In order to reduce processing time and overcome the massive processing of useless frames, Ullah et al. (2019) used a lightweight convolutional neural network (CNN) model to detect people in video streams. Only 16 frame sequences were passed to the 3DCNN model for final prediction, achieving efficient processing. Li et al. (2019b) proposed a model based on 3D convolutional neural networks, utilizing the DenseNet architecture to promote feature reuse and channel interaction, and proposing bottleneck units to further reduce model parameters. Huszar et al. (2023) proposed a method for violence detection based on deep learning. This method uses 3D convolution to model the dynamic relationships between participants and objects, in order to capture the spatial and temporal structure of the data. This method has high computational efficiency, strong practicality, and is suitable for real-world applications.

The above methods are all violence detection tasks based on fully supervised frameworks, and there are also some violence detection methods based on unsupervised frameworks. Ehsan et al. (2022) proposed a double-stream AutoEncoder (Double-AE) violence detection task based on an unsupervised framework. This method is a convolutional autoencoder (AE) implemented by analyzing the potential space of double-stream. Double-AE takes discriminative spatial and temporal information as input samples. Remove background environment and other noise information from video clips in spatial flow. Double-AE extracts the obtained features to lower dimensions to ignore unimportant information and increase time complexity. In order to overcome the problem of insufficient violence data, Ehsan et al. proposed an unsupervised Spatiotemporal Action Translation (STAT) (Ehsan et al., 2024) network to achieve violence detection. The proposed method

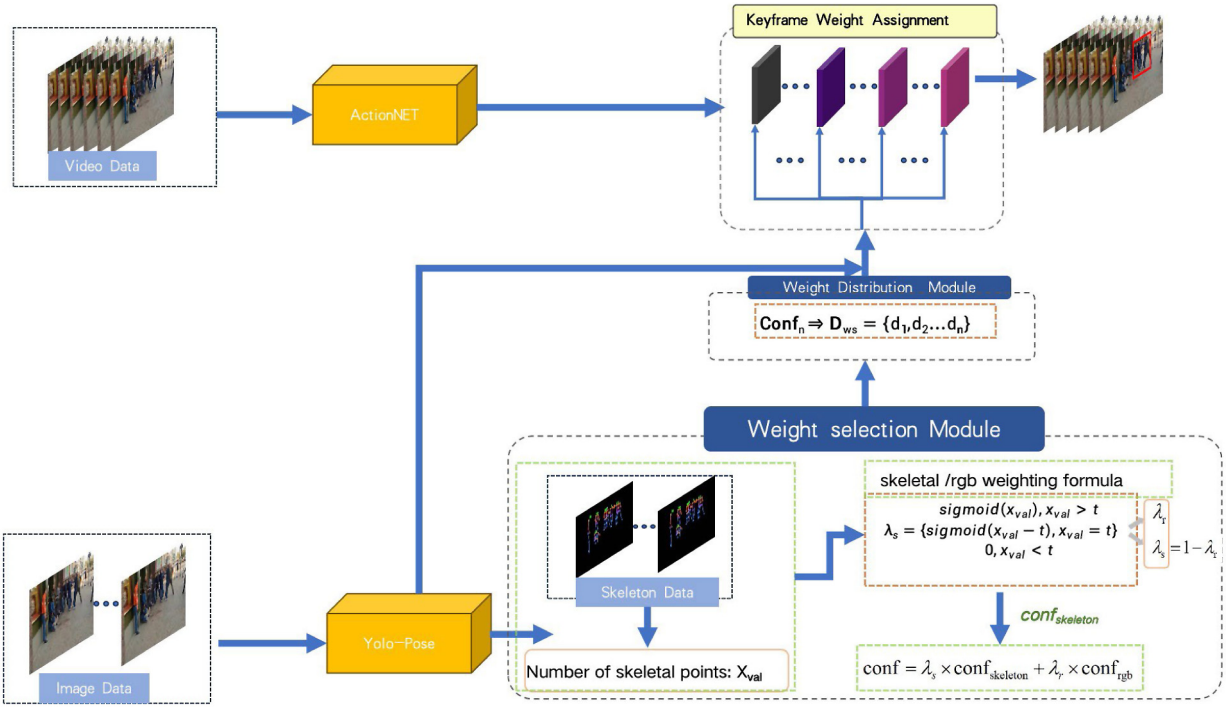



**Fig. 1.** The architecture of RTPNet. The framework of RTPNet can be divided into five parts: backbone network, posture extractor, Weight Selection Module (WSM), Weight Distribution Module (WDM), and Keyframe Weight Assignment (KWA). Firstly, the input data is divided into two types: video clips and images; (1) The backbone network adopts 2DCNN (Action-Net) to ensure the real-time performance of the model; (2) The pose extractor YOLO-Pose takes multiple images as input to extract local skeletal skeleton information and generate skeleton data. (3) Weight Selection Module introduces the parameters generated from skeleton data as a weight coefficient to achieve multimodal interaction between skeletal information and RGB; (4) Weight Distribution Module reduces the error in identifying skeletal information by modifying the confidence level between adjacent frames. (5) Keyframe Weight Assignment (KWA) combines WSM and WDM to output recognized violent behavior results.

is based on adversarial training and can transform temporal features into spatial domains, performing well in different environments. The STAT framework has designed an interpretation section to check for reconstruction errors between actual frames and reconstructed frames, in order to classify human behavior. Compared with Double-AE (Ehsan et al., 2022), the STAT framework has improved model performance, such as precision, recall, and accuracy.

### 2.2. Skeleton-based violent detection

ST-LSTM (Tang et al., 2019) models temporal dynamics of skeleton data based on recurrent neural networks, inputting structural information of skeleton data into LSTM sequence model temporal dynamics. The CNN based method encodes temporal dynamics and skeleton joints into rows and columns respectively, represents skeleton sequences as images, and then passes them onto CNN to recognize their basic actions like image classification. HCN (Li et al., 2018) adopts an end-to-end co-occurrence feature learning framework, which uses CNN to automatically learn hierarchical co-occurrence features from skeleton sequences. The spatiotemporal graph convolutional network (ST-GCN) (Yan et al., 2018) is the first to extend neural networks to graph structures. Design a universal representation of skeleton sequences for action recognition by extending graph neural networks to a spatiotemporal graph model. It considers the human skeleton joints themselves to be a topological graph, representing joint point data as a topological graph, and then inputting the topological graph into a behavior recognition network with a graph convolutional network as the backbone, ultimately obtaining the results of behavior recognition. But graph neural networks are difficult to fuse with other modalities (RGB, FLOW); As the number of people increases, the complexity of GCN increases linearly. There are also some methods (Su et al., 2020, 2023) that process skeleton sequences as input 3D point clouds and learn contextual relationships between relevant personnel from human skeleton points to identify violent behavior. Skeleton based methods cannot accurately capture the local motion relationships of keyframes, for this reason, we propose an effective fusion of motion and appearance information in our method, which focuses on capturing the local motion information of keyframes by enhancing the skeleton key point features between consecutive frames through a unified framework.

## 3. The proposed method

In order to meet the real-time requirements of the algorithm, we have designed a flexible real-time violence detection framework, as shown in Fig. 1. The basic idea of the method is to determine keyframe information by introducing human skeleton feature information, using YOLO-Pose as a pose extractor to construct a human action skeleton map, and combining it with real-time 2DCNN network to classify action videos. We propose a weight selection module and a keyframe weight assignment module to jointly allocate RGB information and skeletal information. This is introduced as inductive bias in our model. The proposed model consists of five modules together; In the following text, we will provide a detailed description of the network architecture along the process and each component.

### 3.1. Human skeleton map extraction

Firstly, this paper constructs a VIMD dataset for extracting human skeleton images. 1500 images from the training set are used for pre training of the YOLO-Pose pose extractor, providing good data conditions for subsequent multi person pose estimation and object key point detection in videos. The experimental results conducted on the VIMD dataset in this paper are shown in 4.1;

Due to YOLO-Pose being a single shot method, it utilizes anchors to associate human joint points. Therefore, it is easy to detect errors when the crowd is dense or when two people are close together. In order

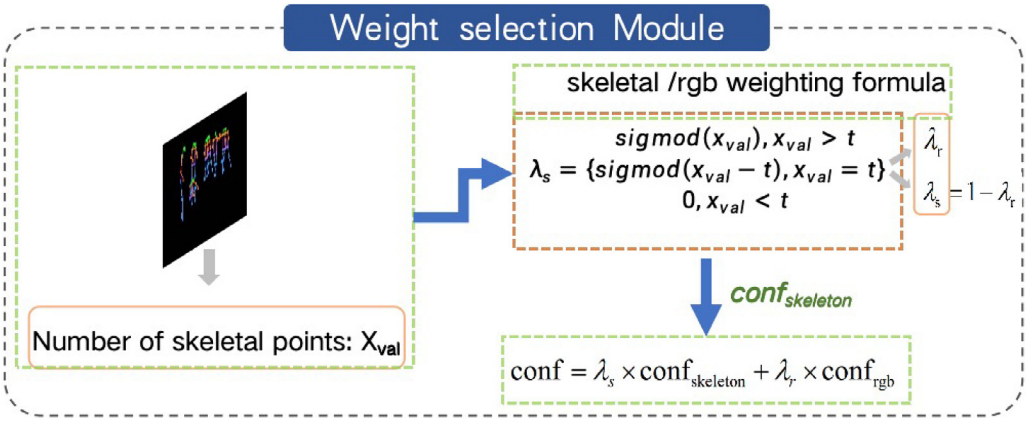



**Fig. 2.** Weight selection module framework diagram.

to obtain more accurate contour points of human joints and objects (collectively referred to as key points). We combine the set of key points extracted from image data into a human skeleton diagram. YOLO-Pose takes the image as input, extracts local skeletal point information, and finally combines it into a skeleton data; Each skeleton map is composed of multiple human key points, storing a sequence of human skeletons. There is only one category, namely human body, with 17 joint points per human body, and each joint point is determined by its position and confidence {x, y, conf}. For each anchor corresponding to a single person with n key points, P can be used $P_v$ represents the predicted vector.

$$P_v = \left\{ C_x, C_y, W, H, box_{conf}, class_{conf}, K_x^1, K_y^1, K_{conf}^1, \ldots\ldots, K_x^n, K_y^n, K_{conf}^n \right\} \tag{1}$$

YOLO-Pose uses CSP-darknet53 as the skeleton, while PANet integrates multi-scale features. Follow the approach of four detection heads at different scales. Finally, each detection head branches into a prediction bounding box and key points. When regressing the position of key points, OKS (Object Keypoint Similarity) is used as the loss function; OKS calculates and sums each key point separately. The announcement is as follows:

$$\mathcal{L}_{kpts}(s,i,j,k) = 1 - \sum_{n=1}^{N_{kpts}} \mathrm{OKS} = 1 - \frac{\sum_{n=1}^{N_{kpts}} \mathrm{EXP}(\frac{d_n^2}{2s^2K_n^2},)\delta(v_n > 0)}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)} \tag{2}$$

Among them, $d_n$ is the Euclidean distance between groundtruth and the predicted point, $K_n$ is the key point weight parameter, and s is the scale of the target, $\delta(v_n)$ represents the visibility of each key point. The output is in two modes: RGB and skeleton. YOLO-Pose takes the image as input, extracts local skeletal point information, and finally combines it into a skeleton data.

### 3.2. 2DCNN network

In order to accommodate the model complexity and inference speed of the framework, we introduce an ACTION module with a mixed attention mechanism for temporal action modeling; The core idea of ACTION-Net (Zhao et al., 2019) is to generate three attention maps, namely; Spatiotemporal excitation of STE paths, channel excitation of CE, and motion excitation of ME to stimulate the corresponding features in the video. The ACTION module is composed of three attention modules in parallel. The ACTION module is based on 2DCNN, which takes video clip as input, and the input is a four-dimensional vector $X \epsilon R^{N \times T, C, H, W}$; $N$ represents the batch size, $T$ represents the number of frames, C represents the number of channels, H represents the height, and W represents the width.

### 3.3. Weight selection module

We propose a weight selection module to effectively distinguish between skeletal information and RGB information in the model; The Weight Selection Module uses the skeleton image pretrained by 3.1 as input, which can be viewed as a sequence of human key points, with each sequence storing the human key points of each image; Each image corresponds to an $X_{val}$, representing the number of skeleton key points. $X_{val}$, as a weight coefficient, is compared with the threshold of limiting skeletal points to determine the weight coefficients of skeletal information confidence $\lambda_s$ and RGB information confidence $\lambda_r$. Thus achieving multimodal interaction between skeleton information and RGB; As shown in Fig. 2, this article designs a weight allocation formula using the sigmoid activation function; Transform the skeleton graph into output on (0,1). The skeleton/RGB weight formula is:

$$conf = \lambda_s \cdot conf_{skeleton} + \lambda_r \cdot conf_{rgb} \tag{3}$$

$$\lambda_s = \begin{Bmatrix} sigmoid\left(x_{val}\right), x_{val} > t \\ sigmoid\left(x_{val} - t\right), x_{val} = t \\ 0, x_{val} < t \end{Bmatrix} \tag{4}$$

$$\lambda_s + \lambda_r = 1 \tag{5}$$

where $X_{val}$ represents the number of skeletal points, $\lambda_s$ is a parameter for assigning confidence in skeletal information, $\lambda_r$ is the parameter for assigning RGB information confidence, $\lambda_s$ and $\lambda_r$ The sum of r is 1, and t is the threshold for limiting skeletal points, which is set to 9 . When the number of skeletal points recognized by the model is greater than or equal to 9, it defaults to no occlusion phenomenon, which makes the model pay more attention to skeletal information; Otherwise, when the number of skeletal points is less than 9, it is considered that the model highlights RGB information, and the parameter for setting the confidence level of skeletal information is 0. The sigmoid function is introduced to normalize the weight coefficients within the [0,1] interval. In short, the module takes the number of skeletal points as the core reference data, and decides whether the skeletal detection branch or the RGB branch should have a higher weight score by the number of skeletal points, because the results of skeletal detection are very accurate. Violence detection based on skeletal maps can effectively remove the background interference of the image and make the detection results more accurate. We will give detailed test results in 4.1.

### 3.4. Weight distribution module

We know that violent action is the moment when there is physical contact is the moment when the highest weight should be assigned. Since the transmission of surveillance video is greater than 24 frames

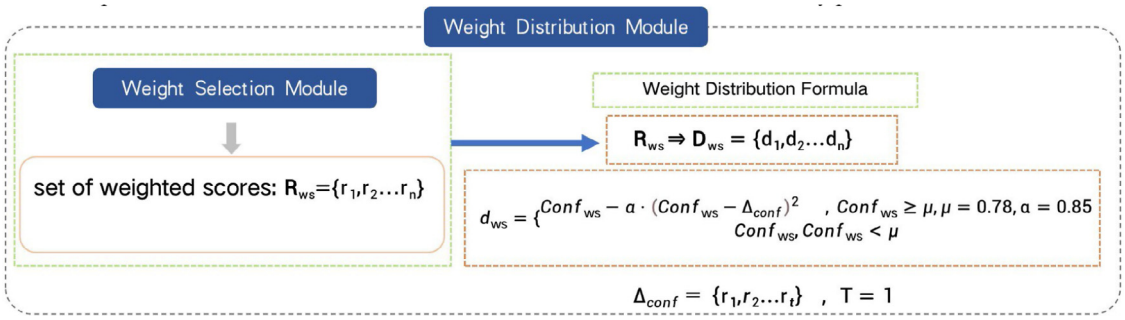



**Fig. 3.** Weight distribution module framework diagram.

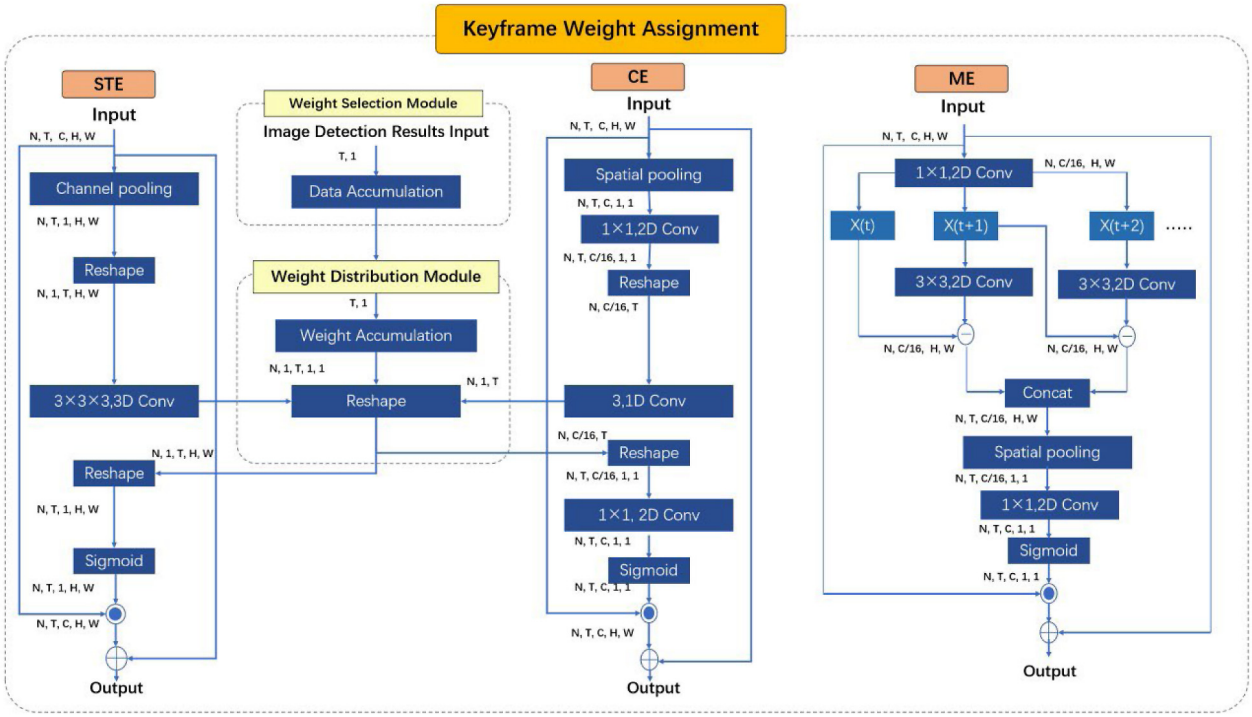



**Fig. 4.** Keyframe weight assignment framework diagram.

per second, the frames before and after the highest moment of physical contact should also have higher weights assigned to them, and their scores should converge to the weights of this moment, but not exceed it. To get a more rational weight distribution, this article proposes a weight distribution module, as shown in Fig. 3, which finds the largest reference frame weight coefficients in the Weight Selection Module(WSM) results and uses the previous and subsequent T-frames as auxiliary reference frames, and its results converge to and are smaller than the reference frame weight results. The skeleton key point confidence set obtained from the weight selection module in 3.3 is denoted by $R_{ws}$ represents, as shown in formula (6), $R_{ws}$ as input to the Weight distribution module, $D_{ws}$ as the output of the Weight distribution module; After co filtering of adjacent $T$ frames, set the allocation weight threshold, and when the confidence of the skeleton point is greater than $\mu$, adaptive adjustment the value of $conf_{ws}$, $\Delta_{conf}$ is the difference between the initial confidence level and the threshold, also known as bias. Bringing it into formula (7) yields $D_{ws}$ as the final weight coefficient.

$$R_{ws} = \{r_1, r_2, \ldots, r_n\}, D_{ws} = \{d_1, d_2, \ldots, d_n\} \tag{6}$$

$$D_{ws} = \begin{cases} conf_{WS} - \alpha * (conf_{WS} - \Delta_{conf})^2, conf_{WS} \geq \mu \\ conf_{WS}, conf_{WS} < \mu \end{cases} \tag{7}$$

$$\Delta_{conf} = \{r_1, r_2, \ldots, r_t\} \tag{8}$$

where $\mu$ Represents the allocation threshold, which is set to 0.78, α Represent weight coefficient, set to 0.85, where T is the input frame number, and this article sets it to 5 frames. $N$ is the number of key points.

### 3.5. Keyframe weight assignment module

As shown in Fig. 4, we propose a keyframe dynamic weight assignment strategy, which is responsible for fusing features from different modalities while capturing feature information on the channel and spatially structured position information. Firstly, based on the data accumulation confidence score output from the Weight Selection Module, it is reshaped into dimensions that can be used for three-dimensional convolution operations, namely (N, 1, T, H, W); Output weight accumulation through the Weight Distribution Module; Secondly, the STE module first convolves the sigmoid to obtain the first spatiotemporal attention map. We fuse and reshape the spatiotemporal features obtained by the STE module with the weight accumulation output from the Weight Distribution Module; The final output of the fused detection results.

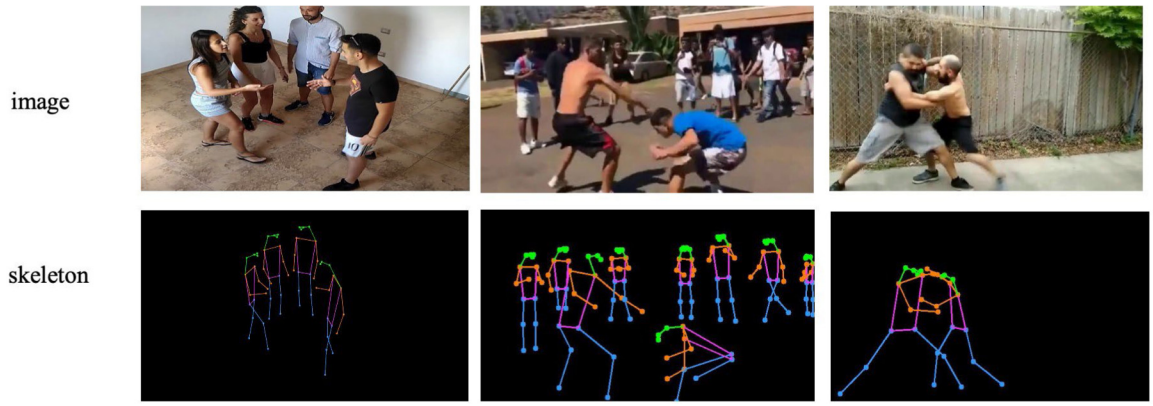



**Fig. 5.** VIMD dataset samples.

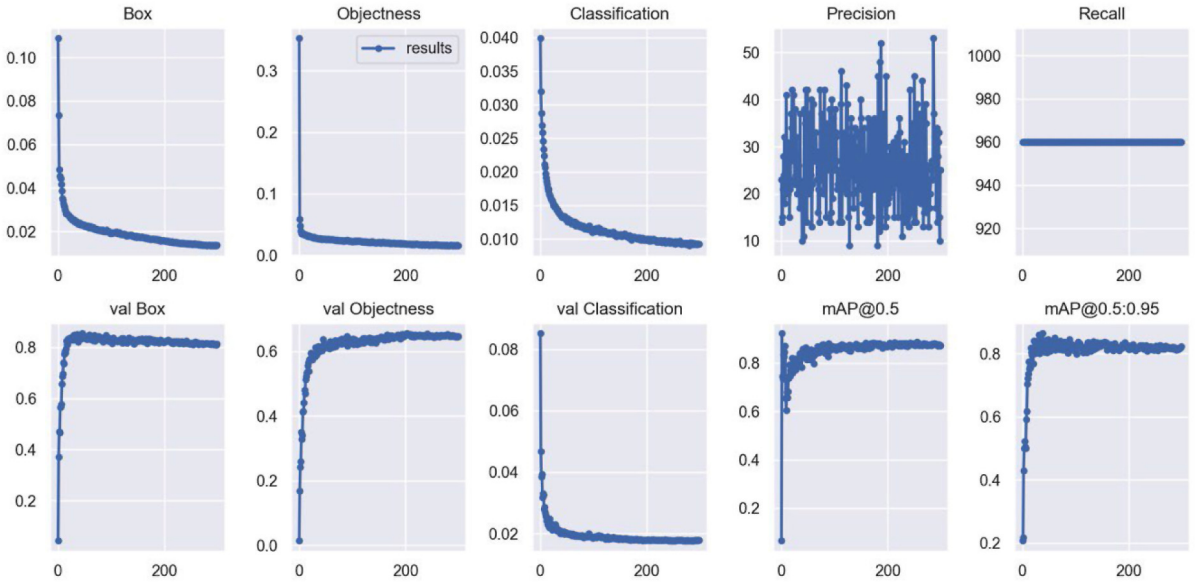



**Fig. 6.** The loss functions results of the VIMD dataset. The horizontal coordinate represents the number of training rounds (epoch = 200).

## 4. Experiments

### 4.1. VIMD dataset

We propose a Violent Image Dataset (VIMD), which consists of various sample images and video frames captured through network collection and monitoring of videos. As shown in Fig. 5, the data was collected from violent behavior images captured from the UBI-Fights (Degardin and Proenca, 2020), AVD (Bianculli et al., 2020) and Hockey-Fight datasets (Nievas et al., 2011), as well as self captured indoor fighting behavior images. Including 2422 fighting and non-fighting images, as well as 1054 labeled skeletal information labels. Given that the main objective is to identify combat scenes on real surveillance camera content, our independent shooting perspective is mainly a top view. The dataset is divided into 1922 training sets and 500 testsets.

As shown in Table 1, comparative experiments were conducted on the image dataset VIMD after introducing skeleton information. After fusing skeleton information, the accuracy of the VIMD dataset was improved to 97.2%, and the recall rate was increased to 98.1%. The results showed that the introduction of skeleton information reduced false positives in the fight action recognition task, and the features extracted through posture increased the accuracy and recall of the model.

**Table 1**
Experimental results of skeleton information on the VIMD dataset.

| | TP | FP | FN | Precision | Recall | FPS |
|---|---|---|---|---|---|---|
| RGB | 400 | 66 | 21 | 85.8% | 95.0% | 36 |
| RGB+POSE | 415 | 12 | 8 | 97.2% | 98.1% | 33 |

The loss functions results of the VIMD dataset are shown in Fig. 6; Box represents the error between the prediction box and the calibration box (CIoU); The confidence loss objectness represents the confidence level of the computational network; Classification loss indicates whether the calculation anchor box and its corresponding calibration classification are correct; mAP@0.5 0.95 represents the average mAP at different IoU thresholds (from 0.5 to 0.95, step size 0.05); mAP@0.5 Represents an average mAP with a threshold greater than 0.5.

As shown in Fig. 7, the last image of each row represents the distribution of x, y, width, and height. The top image (0, 0) indicates the distribution of the horizontal coordinate $x$ of the center point, and it can be seen that most of it is concentrated at the center of the entire

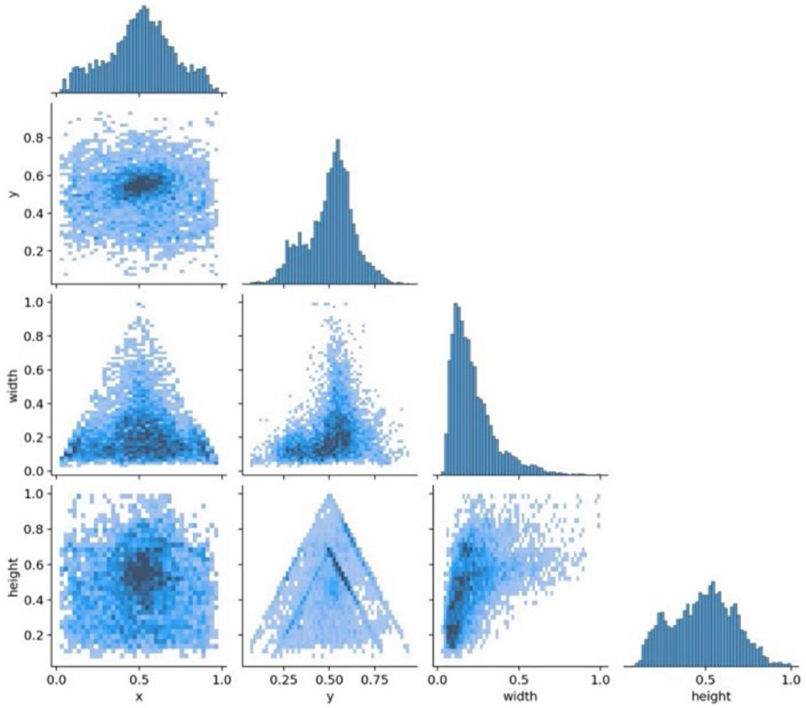



**Fig. 7.** Labels correlogram in VIMD, reflecting the relationship between the horizontal and vertical coordinates of the center point and the height and width of the box.

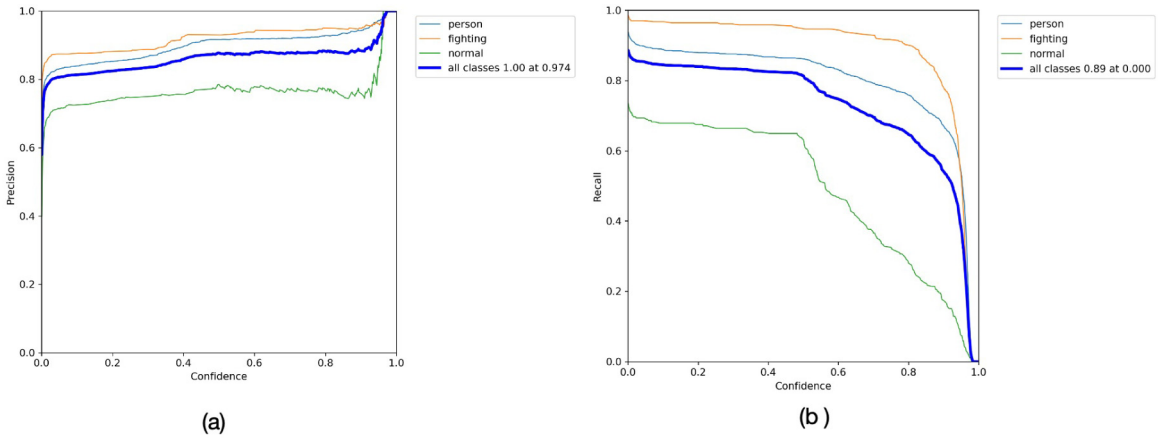



**Fig. 8.** (a) Precision curve; (b) Recall curve.

image; (1,1) The chart shows the distribution of the *y*-axis of the center point, and it can be seen that most of it is concentrated in the center position of the entire map; (2, 2) The chart shows the distribution of box width, and it can be seen that the width of most boxes is about half of the width of the entire image; (3, 3) The chart shows the distribution of box width, and it can be seen that the height of most boxes exceeds half of the height of the entire image; The other graphs are looking for the relationships between these four variables. The precision curve and recall curve are shown in Fig. 8.

### 4.2. Datasets

(1) Hockey-Fight Dataset (Nievas et al., 2011)

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

(2) RWF-2000 (Cheng et al., 2019)

RWF-2000 is a large-scale violent identification dataset collected from YouTube videos. Composed of 2000 video clips captured by surveillance cameras in real scenes. These videos have two actions, either violent or nonviolent, captured by security cameras of various people and backgrounds. There is a 1.6K training and 0.4K testing 5-second video clip, with 30 frames per second. Each video will be annotated with two class tags.

(3) Surveillance Camera Fight (Aktı et al., 2019)

This dataset was collected from YouTube videos containing instances of fights. There are various types of fighting scenes in the dataset, such as kicking, punching, hitting with objects, and wrestling. Due to the different light and shading conditions included in the security camera fragments, these changes were also considered to further increase the diversity of the dataset. In addition, safety camera shots

from different places were collected, such as cafes, bars, streets, buses, shops, etc. The combat scene is independent of the environment of the surveillance camera, and in addition, it includes some fighting sequences from conventional surveillance camera videos. There are a total of 300 videos, 150 fights+150 non fights. There are 120 videos in the training set and 30 videos in the testing set.

(4) AVD (Bianculli et al., 2020)

Automatic Violence Detection in Videos (AVD Dataset) (Bianculli et al., 2020). It consists of 350 segments, labeled as "nonviolent" (120 segments) and "violent" (230 segments) based on the behavior represented. Most available datasets are composed of a few fragments with low resolution, and AVD datasets can be used to test the robustness of brute force detection techniques to false positives. The dataset format is an MP4 video file with a pixel size of 1920 * 1080, video frame rate is 30 fps.

### 4.3. Evaluating indicator

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

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

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

### 4.4. Implementation details

The experiment was trained and tested on four publicly available benchmark datasets. The experimental equipment is an Ubuntu system, with the server configured with Intel (R) Core (TM) i9-9900K CPU, Ge Force RTX 3090Ti, and 32 GB of memory; All code is implemented by Pytorch 1.13.0. The test metric is the accuracy of using each model to detect violent images in the training set. The initialization parameters used in this article's experiment. Use the basic model of resnet50. Input image size is 640 * 640. When the clip length is 32, the batch size is 1; When the clip length is 16, the batch size is 4. The model trained a total of 50 epochs. We use a Random Gradient Descent (SGD) optimizer with momentum and learning rate set to 1e−2.

## 5. Results

To validate the effectiveness of our proposed model, we compared the latest methods on the Hockey-Fight dataset (Nievas et al., 2011), RWF-2000 dataset (Cheng et al., 2019), Surveillance Camera Fight dataset (Aktı et al., 2019), and AVD dataset (Bianculli et al., 2020) in this section. These methods include Skeleton based methods and Appearance based methods. The comparison results of the four datasets are shown in Tables 2 and 3, Table 4, and Table 5, respectively.

### 5.1. Experimental results

The experimental results show that the accuracy obtained on the Hockey-Fight dataset is 99.4%. As shown in Table 2, compare with the appearance based methods, including 3D-CNN (Ding et al., 2014), I3D (Carreira and Zisserman, 2017),FightNet (Zhou et al., 2017), Sudhakaran et al. (Sudhakaran and Lanz, 2017),ECO (Zolfaghari et al., 2018), and TEA (Li et al., 2020). We also compared it with some recent skeleton based methods. Therefore, RTPNet combines these two models and can effectively utilize local and global information. LG-SPIL (Su et al., 2023) is using 3D point cloud technology to extract motion feature information of human skeletal points. By introducing the Skeleton Point Interactive Learning (SPIL (Su et al., 2020)) module, the model has been improved to 97.5%. The model detection accuracy in this article is 99.4%, which is 2.6% higher than the accuracy of the SPIL (Su et al., 2020). The accuracy of RTP-Net is 1.9% higher than the accuracy of the LG-SPIL (Su et al., 2023). As shown in Fig. 9, the training loss and accuracy graph of the Hockey-Fight Dataset at 50 epochs.

**Table 2**
Comparison with state-of-the-arts on the Hockey-Fight dataset.

| Method | Venue (Years) | Video accuracy (%) |
|---|---|---|
| 3D-CNN (Ding et al., 2014) | 2014 | 91.0% |
| I3D (Carreira and Zisserman, 2017) | CVPR2017 | 93.4% |
| FightNet (Zhou et al., 2017) | 2017 | 97.0% |
| Sudhakaran and Lanz (2017) | AVSS2017 | 97.1% |
| ECO (Zolfaghari et al., 2018) | ECCV2018 | 94.0% |
| TEA (Li et al., 2020) | CVPR2020 | 97.1% |
| SPIL (Su et al., 2020) | ECCV2020 | 96.8% |
| Huszár (Huszar et al., 2023) | IEEE2023 | 97.5% |
| LG-SPIL (Su et al., 2023) | CVPR2023 | 97.5% |
| Hachiuma et al. (2023) | CVPR2023 | 99.5% |
| Ours | – | 99.4% |

**Table 3**
Comparison with state-of-the-arts on the RWF-2000 Dataset.

| Method | Venue (Years) | Video accuracy (%) |
|---|---|---|
| C3D (Tran et al., 2015) | ICCV2015 | 77.6% |
| I3D (Carreira and Zisserman, 2017) | CVPR2017 | 83.4% |
| ECO (Zolfaghari et al., 2018) | ECCV2018 | 83.7% |
| DGCNN (Wang et al., 2019) | ACM2019 | 80.6% |
| SPIL (Su et al., 2020) | ECCV2020 | 89.3% |
| TEA (Li et al., 2020) | CVPR2020 | 86.9% |
| LG-SPIL (Su et al., 2023) | CVPR2023 | 90.0% |
| OoD (Sato et al., 2023) | CVPR2023 | 90.3% |
| Huszár (Huszar et al., 2023) | IEEE2023 | 91.0% |
| Hachiuma et al. (2023) | CVPR2023 | 93.4% |
| Ours | – | 93.3% |

In previous studies, the RWF-2000 dataset was used to evaluate the accuracy of violent action recognition in supervised trained models. In this article, the image is first pre-trained based on YOLO-Pose, using violent or nonviolent classification accuracy as an evaluation metric. We tested the accuracy of the proposed method, as shown in Table 3. The highest accuracy comparison between our method and other algorithms is 93.3%. The accuracy is 3.0% higher than OoD (Sato et al., 2023). As shown in Fig. 10, the training loss and accuracy graph of the RWF-2000 Dataset at 50 epochs.

In this article, due to the large number of overlapping scenes and videos in the Surveillance Camera Fight dataset, 300 videos from the dataset and 60 videos from the test set were tested separately. We tested the accuracy of the proposed method, as shown in Table 4. Xception+Bi-LSTM (Aktı et al., 2019) used Xception for spatial modeling and bidirectional LSTM with attention module for temporal modeling. The accuracy comparison between the method proposed in this article and other algorithms reaches a maximum of 93.4%. The accuracy is 1.5% higher than Kang et al. (2021). As shown in Fig. 11, the training loss and accuracy graph of the Surveillance Camera Fight Dataset at 50 epochs.

As shown in Table 5, our model significantly outperformed all methods on the AVD dataset, which validates the superiority of our model. The training set accuracy of the AVD dataset and the validation

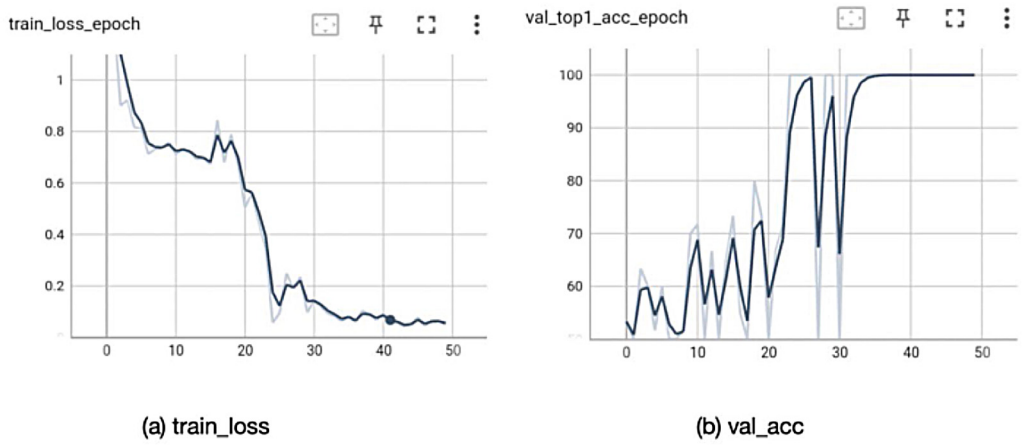



**Fig. 9.** (a) Hockey-Fight Dataset train loss ; (b) Hockey-Fight Dataset validation set accuracy.

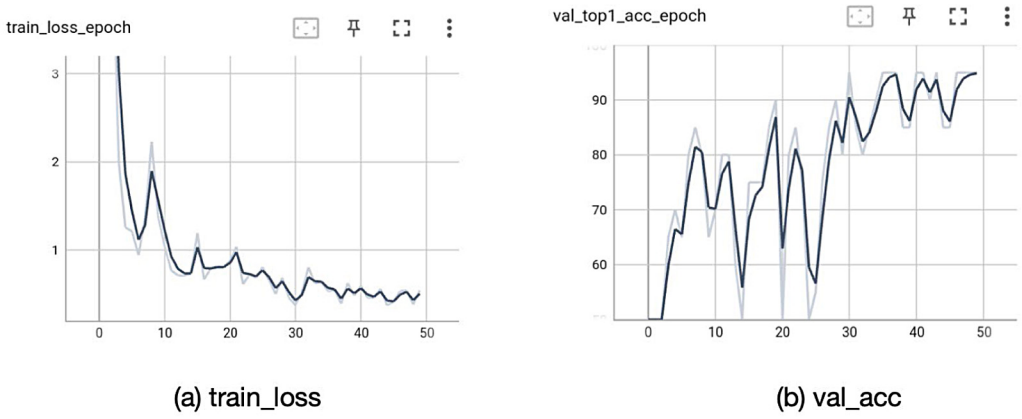



**Fig. 10.** (a) RWF-2000 dataset train loss; (b) RWF-2000 dataset validation set accuracy.

**Table 4**
Comparison with state-of-the-arts on the Surveillance Camera Fight Dataset.

| Method | Venue (Years) | Video accuracy (%) |
|---|---|---|
| Xception+LSTM (Aktı et al., 2019) | IPTA2019 | 55.0% |
| Xception+Bi-LSTM (Aktı et al., 2019) | IPTA2019 | 63.0% |
| Xception + Bi-LSTM+attention (Aktı et al., 2019) | IPTA2019 | 68.0% |
| Fight-CNN + Bi-LSTM (Aktı et al., 2019) | IPTA2019 | 70.0% |
| Fight-CNN + Bi-LSTM+attention (Aktı et al., 2019) | IPTA2019 | 72.0% |
| Kang et al. (2021) | IEEEAccess2021 | 92.0% |
| VD-Net (Ullah et al., 2021) | IEEE2022 | 75.9% |
| ViT (Aktı et al., 2022) | WACV2022 | 84.6% |
| Ours | – | 93.4% |

set accuracy are shown in Fig. 12. The accuracy of our proposed method can reach 99.3%.

### 5.2. Ablation study

In order to explore the optimal parameter settings in this method, we conducted extensive ablation experiments on the RWF-2000 dataset. We set the clip input of the model to 16, 32, and 64 cases; Divide the image size into 640 * 640 and 960 * 960; Improve feature learning for recognition. In order to visually demonstrate the predictive performance of the proposed method, the results show that, as shown in Fig. 13, when the frame number t is 32 and the image size is set to 640 * 640, the accuracy of the model reaches a peak of 93.1%. When the frame number t is 64 and the image size is set to 960 * 960, the accuracy of the model reaches a peak of 93.3%. The longer the video clip, the better the feature extraction ability, and the relatively smaller 640 size as input, the better the performance of the model. when the frame number t is 32, RTPNet has the best real-time performance, with an FPS of 33. As shown in Fig. 14, the visualization results are presented on the RWF-2000 dataset.

**Table 5**
Comparison with state-of-the-arts on the AVD dataset.

| Method | Venue (Years) | Video accuracy (%) |
|---|---|---|
| Sernani et al. (2021) | IEEE2021 | 95.62% |
| Freire-Obregón et al. (2022) | 2022 | 97.54% |
| Ours | – | 99.3% |

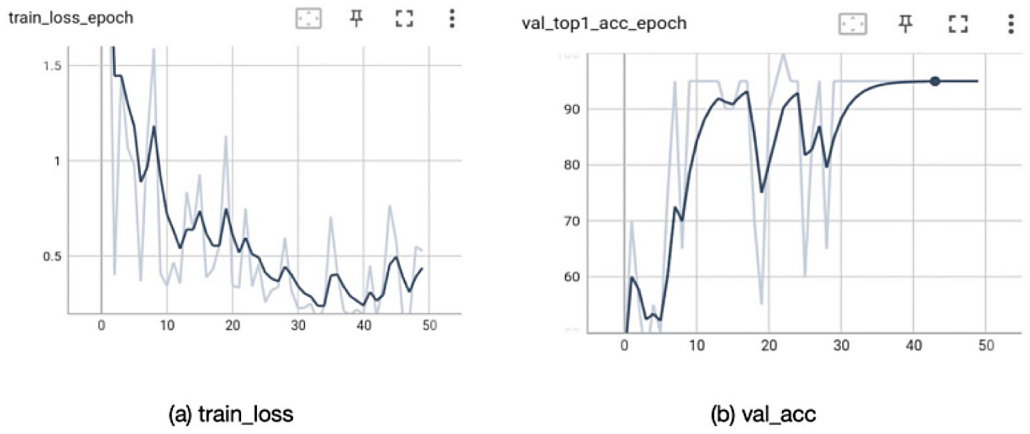



**Fig. 11.** (a) Surveillance Camera Fight dataset train loss; (b) Surveillance Camera Fight dataset validation set accuracy.

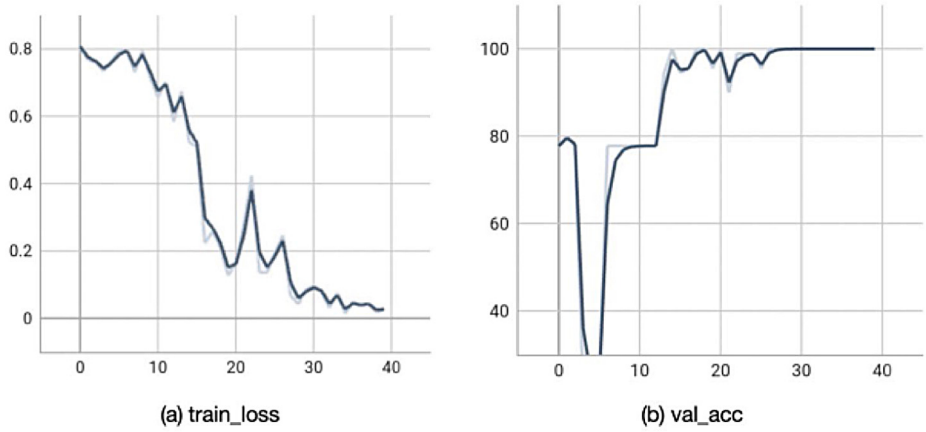



**Fig. 12.** (a) AVD dataset train loss; (b) AVD dataset validation set accuracy.

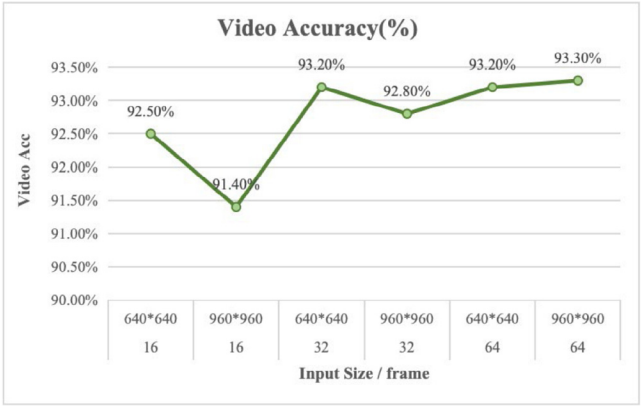



**Fig. 13.** Ablation study on the RWF-2000 Dataset.

Fig. 15 shows the speed comparison of different input sizes on the RWF-2000 dataset, with an input size of 640 * 640 and a clip of 16 being the fastest, reaching up to 33 FPS. Fig. 16 shows the comparison of speed with advanced algorithms. In all three clip scenarios (8, 16, 32), RTPNet has the lowest time consumption and fastest speed, surpassing advanced algorithms such as SlowFast (Feichtenhofer et al., 2019), Timesformer (Bertasius et al., 2021), and uniformerV2 (Li et al., 2022).

In order to validate the effectiveness of the WSM and WDM modules proposed in this paper, ablation study comparison with ACTION-Net is conducted in this paper. As shown in Table 6, the model is optimal on all three datasets such as Hockey-Fight, RWF-2000, and Surveillance Camera Fight dataset. Our proposed algorithm using both WSM and WDM modules on Hockey-Fight achieves 99.4%, which is 5.4% more accurate than ACTION-Net.

## 6. Conclusions

This article proposes a real-time violence detection framework based on keyframe matching. This algorithm uses 2DCNN (ACTION-Net) as the backbone network. This article improves the pre training method of the data. Firstly, a pose extractor named yolo pose is introduced for pre training to extract image level skeleton images as feature information; Secondly, a weight selection module is proposed; The categories generated from stacked human skeleton images serve as a weight coefficient, adaptively selecting which type of biological information the model focuses on (skeletal points, RGB). By default, RGB information is chosen when occlusion is severe; We propose a weight coefficient correction and redistribution module to further modify the coefficients, and Keyframe Weight Assignment (KWA) achieves fusion and interaction between various modules. In addition, in order to verify the effectiveness of introducing skeletal points in extracting fight behavior features, this article created a VIMD image dataset and

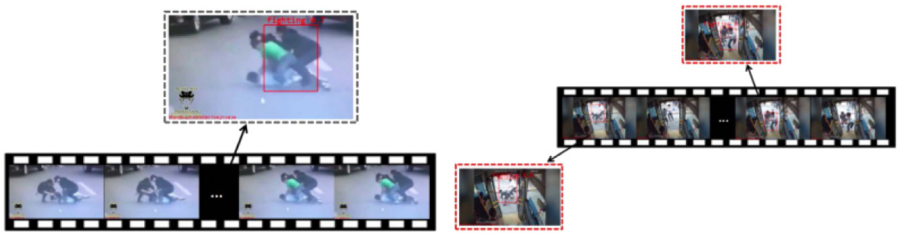


**Fig. 14.** Visualization results on the RWF-2000 dataset.

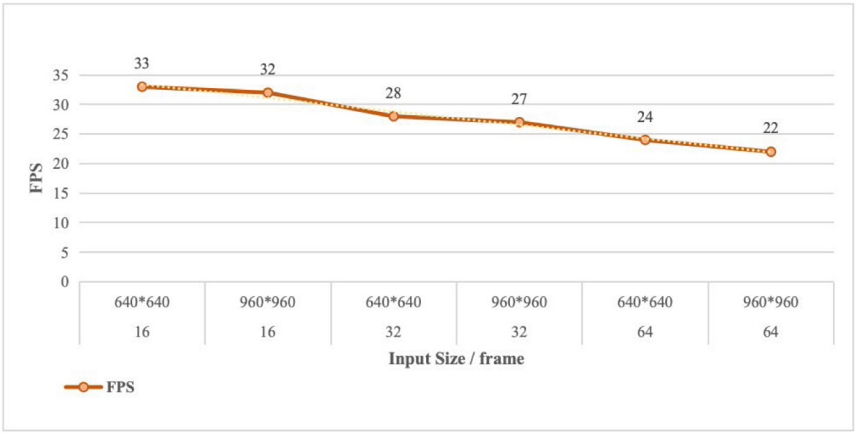



**Fig. 15.** The comparison results of speed on RTPNet.

**Table 6**
Ablation study on Hockey-Fight, RWF-2000, and Surveillance Camera Fight dataset.

| Method | Hockey-Fight(Acc) | RWF-2000(Acc) | Surveillance Camera Fight(Acc) |
|---|---|---|---|
| ACTION-Net (Zhao et al., 2019) | 94.0% | 80.0% | 71.67% |
| ACTION-Net+WSM | 97.95% | 93.28% | – |
| Ours (ACTION-Net+WSM+WDM) | 99.4% | 93.3% | 93.4% |

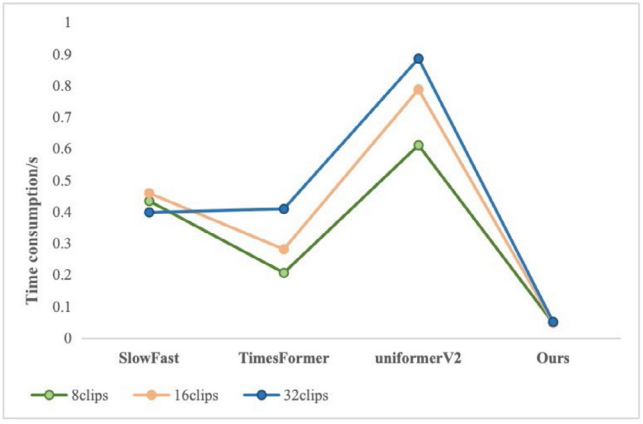



**Fig. 16.** The comparison of speed with advanced algorithms.

conducted experiments on the dataset to compare the detection performance of skeletal information and RGB information in violent scenes. Finally, experiments were conducted on four video level benchmarks, achieving an accuracy of 99.4% on the Hockey-Fight dataset, 93.3% on the RWF-2000 dataset, 93.4% on the Surveillance Camera Fight dataset, and 99.3% on the AVD dataset. The experimental results showed that RTPNet outperformed the SOTA method.

## CRediT authorship contribution statement

**Peng Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xinlei Zhao:** Writing – review & editing, Validation, Methodology, Investigation, Data curation. **Lijia Dong:** Visualization, Validation, Software, Methodology, Conceptualization. **Weimin Lei:** Supervision, Resources, Project administration, Funding acquisition. **Wei Zhang:** Supervision, Software, Project administration, Investigation. **Zhaonan Lin:** Supervision, Project administration, Formal analysis.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

We added the test result files (videos and pictures) to the additional files.

## Acknowledgments

Acknowledgements and Reference heading should be left justified, bold, with the first letter capitalized but have no numbers. Text below continues as normal.

## Appendix. An example appendix

Authors including an appendix section should do so before References section. Multiple appendices should all have headings in the style used above. They will automatically be ordered A, B, C etc.

### *A.1. Example of a sub-heading within an appendix*

There is also the option to include a subheading within the Appendix if you wish.

## References

Aktı, Ş., Ofli, F., Imran, M., Ekenel, H.K., 2022. Fight detection from still images in the wild. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 550–559.

Aktı, Ş., Tataroğlu, G.A., Ekenel, H.K., 2019. Vision-based fight detection from surveillance cameras. In: 2019 9th International Conference on Image Processing Theory, Tools and Applications. IPTA, pp. 1–6.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846.

Bertasius, G., Wang, H., Torresani, L., 2021. Is space–time attention all you need for video understanding? ICML 2 (3), 4.

Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Contardo, P., Lombardi, M., Dragoni, A.F., 2020. A dataset for automatic violence detection in videos. Data Brief 33, 106587.

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.

Chen, L.-H., Hsu, H.-W., Wang, L.-Y., Su, C.-W., 2011. Violence detection in movies. In: 2011 8th International Conference Computer Graphics, Imaging and Visualization. pp. 119–124.

Chen, Z., Li, S., Yang, B., Li, Q., Liu, H., 2021. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1113–1122.

Cheng, M., Cai, K., Li, M., 2019. Rwf-2000: An open large scale video database for violence detection. arXiv preprint arXiv:1911.05913.

Claudi, A., Benedetto, F.Di., Dolcini, G., et al., 2012. Marvin: mobile autonomous robot for video surveillance networks. In: 2012 Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation. IEEE, pp. 21–26.

Degardin, B., Proenca, H., 2020. Human activity analysis: iterative weak/self-supervised learning frameworks for detecting abnormal events. In: 2020 IEEE International Joint Conference on Biometrics. IJCB, IEEE, pp. 1–7.

Demarty, C.-H., Penet, C., Soleymani, M., Gravier, G., 2015. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. Multimedia Tools Appl. 74 (2015), 7379–7404.

Ding, C., Fan, S., Zhu, M., et al., 2014. Violence detection in video by using 3D convolutional neural networks. In: Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December (2014) 8-10, Proceedings, Part II 10. Springer international publishing, pp. 551–558.

Dixit, D.S.K., Dhayagonde, S.B., 2014. Design and implementation of e-surveillance robot for video monitoring and living body detection. Int. J. Sci. Res. Publ. 4 (4), 1–3.

Ehsan, T.Z., Nahvi, M., Mohtavipour, S.M., 2022. Learning deep latent space for unsupervised violence detection. Multimedia Tools Appl. 1–20. http://dx.doi.org/10.1007/s11042-022-13827-7.

Ehsan, T.Z., Nahvi, M., Mohtavipour, S.M., 2024. An accurate violence detection framework using unsupervised spatial–temporal action translation network. Vis. Comput. 40 (3), 1515–1535.

Ergen, T., Kozat, S.S., 2019. Unsupervised anomaly detection with lstm neural networks. IEEE Trans. Neural Netw. Learn. Syst..

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6202–6211.

Freire-Obregón, D., Barra, P., Castrillón-Santana, M., Marsico, M.D., 2022. Inflated 3D ConvNet context analysis for violence detection. Mach. Vis. Appl. 33, 1–13.

Garcia-Cobo, G., SanMiguel, J.C., 2023. Human skeletons and change detection for efficient violence detection in surveillance videos. Comput. Vis. Image Underst. 233, 103739.

Gnouma, M., Ejbali, R., Zaied, M., 2018. Abnormal events' detection in crowded scenes. Multimedia Tools Appl. 77 (2019), 24843–24864.

Hachiuma, R., Sato, F., Sekii, T., 2023. Unified keypoint-based action recognition framework via structured keypoint pooling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22962–22971.

Hassner, T., Itcher, Y., Kliper-Gross, O., 2012. Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–6.

Huszar, V.D., Adhikarla, V.K., Négyesi, I., et al., 2023. Toward fast and accurate violence detection for automated video surveillance applications. IEEE Access 11, 18772–18793.

Islam, Zahidul., Rukonuzzaman, Mohammad., Ahmed, Raiyan., Md. Hasanul Kabir, Farazi, Moshiur., 2021. Efficient two-stream network for violence detection using separable convolutional LSTM. In: IJCNN.

Kang, M.S., Park, R.H., Park, H.M., 2021. Efficient spatio-temporal modeling methods for real-time violence recognition. IEEE Access 9, 76270–76285.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q., 2019a. Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3595–3603.

Li, Y., Ji, B., Shi, X., et al., 2020. Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 909–918.

Li, J., Jiang, X., Sun, T., et al., 2019b. Efficient violence detection using 3d convolutional neural networks. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS, IEEE, pp. 1–8.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y., 2022. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. arXiv preprint arXiv:2211.09552.

Li, C., Zhong, Q., Xie, D., Pu, S., 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv preprint arXiv:1804.06055.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3202–3211.

Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W., 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 143–152.

Maji, D., Nagori, S., Mathew, M., Poddar, D., 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. P. pp. 2637–2646.

Mehran, R., Oyama, A., Shah, M., 2009. Abnormal crowd behavior detection using social force model. In: 2009 IEEEConference on Computer Vision and Pattern Recognition. pp. 935–942.

Mohammadi, S., Kiani, H., Perina, A., Murino, V., 2015. Violence detection in crowded scenes using substantial derivative. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS, pp. 1–6.

Mohtavipour, S.M., Saeidi, M., Arabsorkhi, A., 2022. A multi-stream CNN for deep violence detection in video sequences using handcrafted features. Vis. Comput. 38 (6), 2057–2072.

Moon, G., Kwon, H., Lee, K.M., Cho, M., 2021. Integralaction: Pose-driven feature integration for robust human action recognition in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3339–3348.

Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A.H., Nawaz, H.S., 2021. A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network. In: 2021 6th International Conference on Image, Vision and Computing. ICIVC, pp. 55–60.

Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R., 2011. Violence detection in video using computer vision techniques. In: International Conference on Computer Analysis OfImages and Patterns. pp. 332–339.

Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., 2022. A skeleton-based approach for campus violence detection. Comput. Mater. Contin. 72 (1).

Perez, M., Kot, A.C., Rocha, A., 2019. Detection of real-world fights in surveillance videos. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'19, pp. 2662–2666.

Robert Fisher, J.S.-V., Crowley, James., 2004. CAVIAR: Context aware vision using image-based active recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

Sato, F., Hachiuma, R., Sekii, T., 2023. Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6471–6480.

Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P., Dragoni, A.F., 2021. Deep learning for automatic violence detection: Tests on the AIRTLab Dataset. IEEE Access 9, 160580–160595.

Su, Y., Lin, G., Wu, Q., 2023. Improving video violence recognition with human interaction learning on 3D skeleton point clouds. arXiv preprint arXiv:2308.13866.

Su, Y., Lin, G., Zhu, J., Wu, Q., 2020. Human interaction learning on 3d skeleton point clouds for video violence recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August (2020) 23–28, Proceedings, Part IV 16. Springer International Publishing, pp. 74–90.

Sudhakaran, S., Lanz, O., 2017. Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS, IEEE, pp. 1–6.

Sultani, W., Chen, C., Shah, M., 2018. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6479–6488.

Tang, Q., Yang, M., Yang, Y., 2019. ST-LSTM: A deep learning approach combined spatio-temporal features for short-term forecast in rail transit. J. Adv. Transp. 1–8.

Tran, D., Bourdev, L., Fergus, R., et al., 2015. Learning spatio temporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.

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.

Ullah, F.U.M., Muhammad, K., Haq, I.U., Khan, N., Heidari, A.A., Baik, S.W., de Albuquerque, V.H.C., 2021. AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. IEEE Trans. Ind. Inform. 18 (8), 5359–5370.

Ullah, F.U.M., Ullah, A., Muhammad, K., et al., 2019. Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors 19 (11), 2472.

Vosta, S., Yow, K.-C., 2022. A CNN-?RNN combined structure for real-world violence detection in surveillance cameras. Appl. Sci. 12 (2022), 1021.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. (tog) 38 (5), 1–12.

Yan, S., Xiong, Y., Lin, D., 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32, No. 1.

Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N., 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. P. pp. 1112–1121.

Zhao, D., Xing, Z., Chen, C., Xia, X., Li, G., 2019. ActionNet: Vision-based workflow action recognition from programming screencasts. In: 2019 IEEE/ACM 41st International Conference on Software Engineering. ICSE, IEEE, pp. 350–361.

Zhou, P., Ding, Q., Luo, H., 2017. Violent interaction detection in video based on deep learning. In: Journal of Physics: Conference Series. Vol. 844, IOP Publishing, 012044, (1).

Zolfaghari, M., Singh, K., Brox, T., 2018. Eco: Efficient convolutional network for online video understanding. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 695–712.