Electronic Research Archive

ERA, 32(2): 874–896.
DOI: 10.3934/era.2024042
Received: 16 November 2023
Revised: 15 December 2023
Accepted: 02 January 2024
Published: 15 January 2024

http://www.aimspress.com/journal/ERA

Research article

A novel approach for enhanced abnormal action recognition via coarse and precise detection stage

Yongsheng Lei[1], Meng Ding[1,2,*], Tianliang Lu[3], Juhao Li[1], Dongyue Zhao[1] and Fushi Chen[1]

[1] School of Criminal Investigation, People's Public Security University of China, Beijing 100038, China
[2] Public Security Behavioral Science Lab, People's Public Security University of China, Beijing 100038, China
[3] School of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China

* **Correspondence**: Email: dingmeng@ppsuc.edu.cn.

Abstract: With the proliferation of urban video surveillance systems, the abundance of surveillance video data has emerged as a pivotal asset for enhancing public safety. Within these video archives, the identification of abnormal human actions carries profound implications for security incidents. Nevertheless, existing surveillance systems primarily rely on conventional algorithms, leading to both missed incidents and false alarms. To address the challenge of automating multi-object surveillance video analysis, this study introduces a comprehensive method for the detection and recognition of multi-object abnormal actions. This study comprises a two-stage framework: the coarse detection stage employs an enhanced YOWOv2E model for spatio-temporal action detection, while the precise detection stage utilizes a two-stream network for precise action classification. In parallel, this paper presents the PSA-Dataset to address the current limitations in the field of abnormal action detection. Experimental results, collected from both public datasets and a self-built dataset, illustrate the effectiveness of the proposed method in identifying a wide spectrum of abnormal actions. This work offers valuable insights for automating the analysis of human actions in videos pertaining to public security.

Keywords: deep learning; abnormal action detection; digital forensics; public security; self-attention

1. Introduction

Surveillance videos are integral to maintaining public safety and upholding social order by serving as valuable sources of forensic evidence. However, the current state of intelligent surveillance systems, while capable of real-time detection of abnormal video actions [1], is predominantly characterized by intricate, multi-stage pipelines. These pipelines involve processes like object detection, tracking, identification, and action analysis. Unfortunately, these approaches heavily rely on handcrafted features and struggle to adapt to the diverse landscape of monitoring data.

Video content analysis is indispensable in digital forensics [2], as human motions contain vital biological characteristics for suspect action analysis and case resolution. This research aims to introduce an end-to-end, trainable deep learning approach to automatically detect and precisely recognize abnormal actions in multi-object surveillance videos. This development is poised to empower public safety agencies, granting them the ability to perform efficient forensic video analysis.

This paper confronts the challenge of detecting abnormal actions in surveillance videos, a domain often characterized by multiple objects exhibiting diverse action categories. Traditional action detection methods fall into two categories: temporal and spatio-temporal. Temporal methods focus on identifying actions and their temporal boundaries, while spatio-temporal methods incorporate spatial information into the analysis. This study delves into the domain of spatio-temporal action detection methods, with a specific emphasis on enhancing the effectiveness of action detection in the context of abnormal actions.

Numerous studies have delved into abnormal action recognition in surveillance, largely depending on action detection to identify actions. However, the complexity of action detection is not merely about identifying actions; it is also about precisely localizing them. This intricacy makes training more demanding and balancing recognition and localization accuracy a formidable task. In light of this challenge, this paper advocates for the division of abnormal action recognition in surveillance into two distinctive stages: coarse and precise detection. For untrimmed surveillance videos, the initial stage involves conducting coarse detection using spatio-temporal action detection to identify action categories and temporal boundaries of each object, consequently creating action tubes. Subsequently, precise detection comes into play, utilizing action recognition methods with higher classification accuracy to classify the action tubes.

This study introduces an innovative approach for spatio-temporal action detection and recognition. This work integrates the advanced YOWOv2E model with the ViTSN model, enhancing the precision of multi-object abnormal action recognition while maintaining a rapid inference speed. In response to the practical requirements of public security applications, this research also presents a novel surveillance action dataset, ensuring adaptability of the model to a wide range of real-world scenarios.

The contributions of this paper can be succinctly summarized as follows:

1) In the coarse detection phase, this paper introduces the YOWOv2E algorithm, which builds upon YOWOv2 by incorporating a channel attention mechanism and a novel joint loss function to enhance coarse detection accuracy while maintaining speed.
2) For precise detection, this paper proposes a spatio-temporal two-stream network model based on the Vision Transformer (ViT), utilizing transfer learning to mitigate overfitting and incorporating the Simple Attention Mechanism (SimAM) to reduce background interference. Segmented sampling effectively manages lengthy time sequences, while the integration of optical and RGB data enhances model accuracy.

3) This paper creates a public security surveillance abnormal action dataset (PSA-Dataset) to meet practical forensic demands and validate the generalization performance of the proposed model using this dataset.

2. Related work

This paper divides the process of recognizing abnormal actions into two distinct tasks: action detection and action recognition. Action detection primarily involves the identification and localization of human actions in untrimmed videos, while action recognition focuses on classifying these actions. This research explores the challenges and methods associated with this dual objective.

Action detection: Action detection seeks to identify and localize specific actions within videos. Early methods [3,4] heavily relied on manually crafted features and simplistic classifiers, limiting their effectiveness due to the absence of high-level semantic understanding of complex dynamic actions. The introduction of Convolutional Neural Networks (CNNs) revolutionized video action detection with deep learning [5,6]. Further advancements, such as 3DCNNs [7,8], improved the ability of the model to extract temporal characteristics. In the context of multi-object scenarios, existing methods often entail object detection followed by action localization and classification, posing challenges in end-to-end training, processing efficiency, and deployment. The YOWO series [9,10] introduced a groundbreaking approach that enables simultaneous localization of action boundaries, action recognition, and actor identification. This approach facilitates end-to-end training while addressing timeliness concerns in action detection. One of the publicly available datasets for this assignment is UCF101-24, which offers rich scenarios for behavior analysis and covers a broad spectrum of behavioral categories. JHMDB-21 facilitates more in-depth behavior analysis by providing comprehensive annotations of joint positions in every video. The AVA dataset encompasses a wide range of interpersonal behavior exchanges and is sourced from movie and TV program clips. The UCF-Crime and RWF-2000, sourced from surveillance videos, include real-world anomalous behavior events, contributing to the development and testing of anomaly detection systems applicable in real-world settings.

Action recognition: The primary goal of action recognition is to accurately classify pretrimmed video actions. Initially, CNNs [11] extracted spatial features directly from video frames for classification but did not surpass traditional methods like improved dense trajectories (iDT) in recognition accuracy. The optical method, which calculates object motion information between frames, proved effective in extracting temporal characteristics. The two-stream network [12], combining video frames and stacked optical maps for spatial and temporal feature extraction, enhanced temporal feature extraction. To capture long-term features, Temporal Segment Networks (TSN) [13] introduced a sparse sampling strategy. Recent developments in this field focus on the Transformer architecture [14–17], renowned for its self-attention mechanism that can gather global information directly in both spatial and temporal dimensions. Nevertheless, challenges in action recognition persist, including the need for extensive annotated data, training with small datasets, managing background interference, and addressing computational constraints when processing lengthy time sequences. UCF101, which covers a wide range of action categories from sports to musical instrument performance to daily activities, is one of the public datasets used for this job. The HMDB-51, which spans a wide spectrum of human motions from basic gestures to intricate movements, is gathered from YouTube and motion pictures. Large-scale datasets like Kinetics make it easier to train more intricate and effective models.

3. Abnormal action recognition

3.1. Framework of abnormal action recognition

This paper introduces a two-step approach to digital forensics, encompassing coarse detection and precise detection, to facilitate the identification and detection of abnormal actions in surveillance videos. The overall detection framework is graphically depicted in Figure 1. In the coarse detection phase, the YOWOv2E algorithm builds upon YOWOv2 by introducing a channel attention mechanism and a novel joint loss function, enhancing coarse detection accuracy while maintaining speed. For precise detection, the paper proposes a spatio-temporal two-stream network model based on the Vision Transformer (ViT) [18]. Transfer learning is employed to minimize overfitting, the SimAM [19] reduces background interference, and segmented sampling manages lengthy time sequences. Integrating optical and RGB data within the two-stream network, coupled with the self-attention mechanism, reduces computational complexity and improves accuracy. Constructed action tubes are subsequently classified to ensure robust detection accuracy. To address practical forensic requirements, a public security surveillance abnormal action dataset has been created and employed to validate the performance of the model.

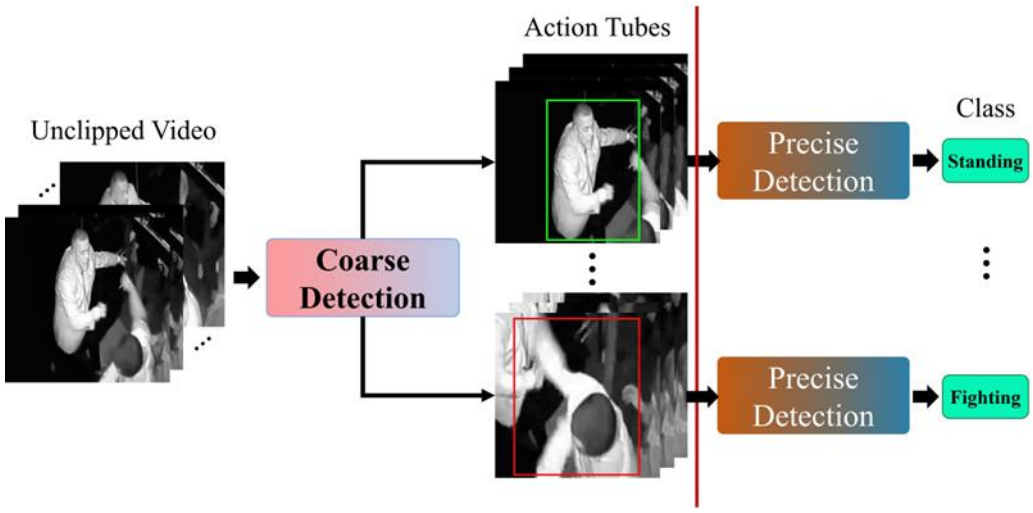

Figure 1. The framework of multi-target abnormal action detection.

3.2. Coarse detection stage

This research introduces the YOWOv2E model, as depicted in Figure 2. At its core, YOWOv2E features two enhanced single-stage networks, 2DShuffleNetv2E and 3DShuffleNetv2E, each incorporating dual branches. These branches are dedicated to the extraction of 2D spatial information and 3D spatio-temporal information, respectively. A fusion network integrates these components through a sequence of encodings, generates detection boxes for human actions. These boxes include confidence scores and categories, enabling precise localization and recognition of actions.

In the YOWOv2E model, the 2D branch focuses on capturing spatial details within video frames, encompassing the shapes, appearances, positions, and postures of individuals. This branch utilizes 2D

convolutional networks, drawing inspiration from the YOLOv7 [20] model in object detection. YOWOv2E treats object detection as a regression problem, directly predicting bounding boxes and action class probabilities from image pixels.

In contrast, the 3D branch is responsible for capturing spatio-temporal information, including action types, speeds, and individual motion directions. This branch predominantly utilizes 3D convolutional networks. The design of the 3D branch equips the YOWOv2E model to comprehensively extract temporal features from actions, enhancing its understanding of actions within videos. YOWOv2E effectively integrates spatial and temporal features, leveraging the ECA [22] module to enhance its feature representation capability.

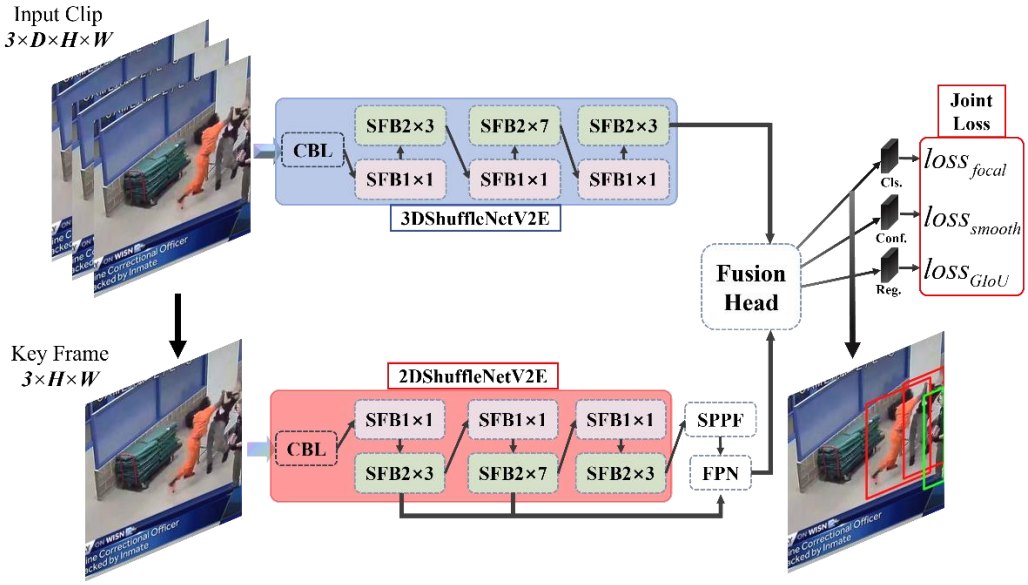

Figure 2. The structure of the improved YOWOv2.

3.2.1. Feature extraction network

The PSA-Dataset, designed for surveillance scenarios, presents unique challenges characterized by complex multi-person interactions and a diverse array of action categories. To better capture features within these intricate scenes, this paper introduces an enhanced ShuffleNetv2 structure, referred to as ShuffleNetv2E. The original ShuffleNetv2 [21] is an efficient network tailored for mobile devices, employing techniques such as channel shuffling and group convolution to enhance computational efficiency. To further optimize its performance in complex surveillance scenarios, this work integrates an Efficient Channel Attention (ECA) [22] attention mechanism at the conclusion of each branch within every Shuffle Block in the ShuffleNetv2 architecture.

The ECA attention mechanism aids in improving the model's focus on the channel elements that are most beneficial for categorization. The features travel through a one-dimensional convolution layer after undergoing global average pooling as input into the ECA attention mechanism. The kernel size of this layer is adaptively adjusted according to the feature map's dimensions. This approach effectively captures the relationships between channels without explicitly increasing the model's parameters or computational complexity. Channel attention weights for the original feature map are

produced by passing the features through a Sigmoid function following convolution layer processing. These channel attention weights lessen the influence of irrelevant features and increase the focus on useful features for anomaly identification. Adaptive weighting across channels is accomplished through element-wise multiplication of these weights with the original features. The model's capacity to describe certain category features is improved by highlighting the dependencies between channels, which further increases the accuracy of anomalous behavior identification.

ShuffleNetv2E centers around the Shuffle Block as its core component, with its operation contingent on the stride value. In the case of a stride value set to 1, as illustrated in Figure 3(a), the input is divided into two components. One portion forms a "direct connection" or "shortcut" to maintain information continuity and mitigate gradient vanishing issues. The other undergoes a 1 × 1 convolution for channel transformation, followed by a 3 × 3 depthwise separable convolution, extracting spatial information while preserving computational efficiency. This is followed by another 1 × 1 convolution, coupled with the ECA mechanism. The result is then concatenated with the "direct connection" and subjected to a channel shuffle operation, facilitating information exchange between features and yielding the final output.

In the case of a Shuffle Block with a stride of 2, as depicted in Figure 3(b), the input is initially bifurcated into two components. The first component undergoes a 3 × 3 depthwise separable convolution with a stride of 2, reducing spatial dimensions while capturing spatial information. It is followed by a 1 × 1 convolution for channel transformation. Concurrently, the second component commences with a 1 × 1 convolution for channel transformation, followed by a 3 × 3 depthwise separable convolution with a stride of 2 to reduce spatial dimensions. This part also undergoes another 1 × 1 convolution for channel transformation and benefits from the ECA mechanism to enhance feature distinctiveness.

Following the described operations, the two separate feature map parts are combined. To promote effective information exchange in the channel dimension between these components, they are subjected to a channel shuffle operation. This process results in the output of the Shuffle Block, and this outcome varies depending on whether the stride is set to 2 or 1. Following these operations, the two feature map components are unified. To facilitate effective information exchange in the channel dimension between these components, they are subjected to a channel shuffle operation. This process culminates in the output of the Shuffle Block, and the resulting output varies based on whether the stride is set to 2 or 1.

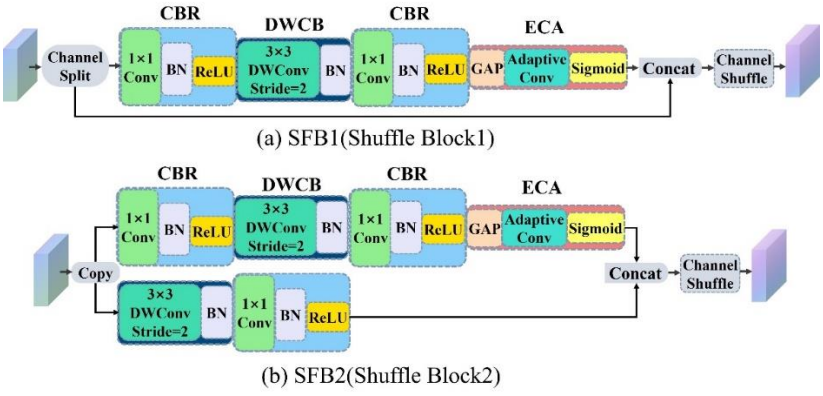

Figure 3. The structure of the improved Shuffle Block. (a) The structure of the Shuffle Block 1. (b) The structure of the Shuffle Block 2.

3.2.2. Joint loss function

In the context of modeling abnormal actions, a common and challenging problem arises due to class imbalance. Abnormal events are less frequent, resulting in a dataset with significantly more normal action samples than abnormal action samples. This imbalance can introduce bias, where the model is more likely to favor the majority class.

In the context of modeling abnormal actions, the challenge of class imbalance frequently arises. Abnormal events are less frequent, leading to a dataset with significantly more normal samples than abnormal ones. This imbalance can introduce bias, with the model favoring the majority class. To effectively address this issue, this paper employs the Sigmoid Focal Loss [23] as the classification loss function, defined in Eq (1). In this equation, p_t represents the predicted probability for positive samples, α denotes the weight coefficient, and γ regulates the weight distribution. Specifically designed for scenarios where each category is treated as an independent binary classification problem, especially in cases of multiple categories or multi-label tasks, the Sigmoid Focal Loss assigns lower weights to normal actions and higher weights to abnormal actions. This encourages the model to prioritize the learning of abnormal actions during training.

$$L_{focal}(p_t) = -\alpha(1-p_t)^{\gamma}\ln(p_t) \tag{1}$$

When conducting a detailed analysis of abnormal actions, a common challenge lies in the ambiguity of action boundaries. Many actions, such as walking and running or sitting and falling, lack clearcut boundaries, making it challenging for the model to provide accurate predictions, particularly at these ambiguous boundaries. To address this challenge and enhance the resilience of the proposed model in handling these fuzzy boundaries, this paper introduces Label Smooth [24] into the Binary Cross-Entropy (BCE) with Logits Loss, utilizing it as a confidence loss function. Label Smooth transforms predictions from hard labels (e.g., 0 or 1) into softer labels ranging between 0 and 1, as shown in Eqs (2) and (3). By introducing a small smoothing parameter ε, Label Smooth offers a more gradual learning approach, preventing the model from making overly confident predictions. Instead, it adds flexibility to the predictions of the model, contributing to better handling of boundary samples among various actions.

$$\hat{y} = (1-\varepsilon)y + \frac{\varepsilon}{2} \tag{2}$$

$$L_{smooth}(\hat{y}, p) = -\hat{y}\ln(p) - (1-\hat{y})\ln(1-p) \tag{3}$$

In this paper, the training loss function, as depicted in Eq (4), combines three components: L_{conf} (BCE with Logits Loss with Label Smooth integration), L_{cls} (Sigmoid Focal Loss), and L_{reg} (GIoU Loss). This combination aims to enhance the model's overall performance. Within the equation, $a_{x,y}$, $b_{x,y}$, and $c_{x,y}$ correspond to classification predictions, regression predictions, and confidence predictions, respectively, while $\hat{a}_{x,y}$, $\hat{b}_{x,y}$, and $\hat{c}_{x,y}$ represent the corresponding labels. N_{pos} signifies the number of positive samples, and α, β, and θ are the weights assigned to each loss component. $I_{\{\hat{a}_{x,y}>0\}}$ serves as an indicator function that evaluates to 1 when the condition $\hat{a}_{x,y} > 0$ holds and 0 otherwise.

$$L(\{a_{x,y}\},\{b_{x,y}\},\{c_{x,y}\}) = \frac{\theta}{N_{pos}}\sum_{x,y} L_{conf}(\hat{c}_{x,y}, c_{x,y}) + \frac{\alpha}{N_{pos}}\sum_{x,y} I_{\{\hat{a}_{x,y}>0\}} L_{cls}(\hat{a}_{x,y}, a_{x,y}) + \frac{\beta}{N_{pos}}\sum_{x,y} I_{\{\hat{a}_{x,y}>0\}} L_{reg}(\hat{b}_{x,y}, b_{x,y}) \tag{4}$$

3.3. Precise detection stage

Precise detection is carried out to reduce the false alarm rate of the algorithm after constructing action tubes from untrimmed surveillance videos. The enhancement of action detection in surveillance videos leverages the ViTSN model, composed of four key modules: an attention module based on SimAM, spatio-temporal feature extraction using the pretrained ViT, a time self-attention module, and a decision fusion module. Furthermore, improvements have been made to the loss function. The process involves preprocessing action tubes to obtain optical frames, feature extraction with the SimAM-based attention module and pretrained ViT, extracting temporal information with the temporal attention module, and assigning weights to RGB and optical images using the decision fusion module. The significance and role of each module in reducing false alarms and increasing detection accuracy are detailed in Figure 4 of the source.

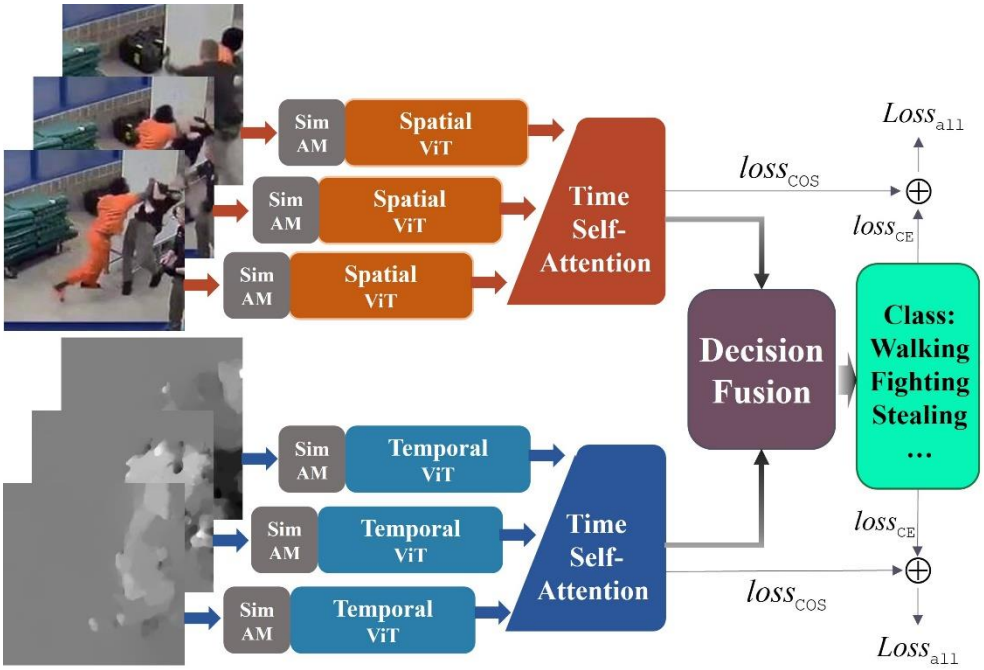

Figure 4. Spatio-temporal two-stream network model based on ViT.

3.3.1. Simple attention mechanism

The primary innovation of the model lies in the introduction of a 3D parameter-free attention

mechanism into the backbone network to enhance interference resistance. This mechanism focuses on information-rich neurons and their impact on neighboring neurons, particularly regarding spatial suppression effects. It defines an energy function for each neuron, quantifying the linear separability between the target neuron and others. The importance of individual neurons is assessed, and their weights are adjusted accordingly, as depicted in Eq (5). Equation (5) represents the energy function, where t denotes the target neuron, $e_t{}^*$ signifies the energy of the target neuron, $\hat{\mu}$ and $\hat{\sigma}^2$ represent the mean and variance of all other neurons in a single channel, excluding the specified neuron, and λ is a hyperparameter. Explain the practical implications and benefits of this approach.

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{\left(t - \hat{\mu}\right)^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{5}$$

Equation (6) implements the assessment of neurons through a linearly separable function, where E represents a matrix composed of $e_t{}^*$, $sigmoid$ is the activation function, which introduces nonlinearity. The importance of a target neuron is inversely related to its energy level, allowing it to distinguish itself from neighboring neurons and emphasizing its significance. Subsequently, this target neuron's weight is determined based on its significance. Clarify the implications and advantages of this evaluation approach within the context of neural network models.

$$\tilde{\boldsymbol{X}} = \texttt{sigmoid}\left(\frac{1}{\boldsymbol{E}}\right) \odot \boldsymbol{X} \tag{6}$$

In contrast to the one-dimensional squeeze-and-excitation network (SENet) and the two-dimensional convolutional attention mechanism (CBAM), the three-dimensional SimAM is proficient in assessing the significance of all neurons. This distinctive capability enables SimAM to focus on both the channel and spatial dimensions. This streamlined approach not only diminishes background interference but also augments action features.

3.3.2. Self-attention mechanism

Both two-stream feature extraction networks are built upon the ViT model but use distinct input data sources. The spatial network takes a sequence of n RGB image frames as input, facilitating action classification through the extraction of static image features, involving feature fusion from multiple frames along the temporal dimension. In contrast, the temporal network is fed n consecutive sets of optical images, each comprising 10 frames, which encode motion information, contributing to the extraction of action-related data.

This paper introduces a temporal self-attention mechanism, enabling the model to detect changes in temporal features and enhance their extraction. The spatial features extracted by ViT are reshaped into a matrix and undergo temporal self-attention operations. This reshaped matrix is subjected to linear projection, yielding the query matrix $\boldsymbol{Q}$, the key matrix $\boldsymbol{K}$, and the value matrix $\boldsymbol{V}$. The complete temporal self-attention process is illustrated in Eq (7).

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = SoftMax(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}) \cdot \boldsymbol{V} \tag{7}$$

The complete temporal self-attention process is illustrated in Eq (8) for the input matrix. In this

equation, $LayerNorm()$represents a normalization layer designed to stabilize the feature distribution, $MLP()$denotes a multi-layer perceptron responsible for scaling feature vectors to extract more abstract and meaningful feature representations, and $Dropout()$is a regularization technique that selectively deactivates certain neurons to mitigate the risk of model overfitting.

$$\begin{aligned} \boldsymbol{X}_1 &= Attention(LayerNorm(\boldsymbol{X})) \\ \boldsymbol{X}_2 &= \boldsymbol{X}_1 + Dropout(\boldsymbol{X}_1) \\ \boldsymbol{X}_3 &= MLP(LayerNorm(\boldsymbol{X}_2)) \\ Output &= \boldsymbol{X}_3 + Dropout(\boldsymbol{X}_3) \end{aligned} \tag{8}$$

3.3.3. Joint loss function

The loss function of the network model under consideration places a premium on enhancing classification precision while concurrently minimizing feature disparities across a multitude of video segments. The formula for these calculations is articulated as follows:

$$\begin{aligned} loss_{\text{CE}} &= -\sum_{i=1}^{n} y_i \ln y_i \\ loss_{\cos} &= \max\left[1-\cos(x_i, x_j)\right] \\ Loss_{\text{all}} &= \lambda loss_{\text{CE}} + \eta loss_{\cos} \end{aligned} \tag{9}$$

The loss function of the network model prioritizes enhancing classification precision while minimizing feature disparities across numerous video segments. The formula for these calculations, as shown in Eq (9), incorporates the cross-entropy loss function, which quantifies the disparity between the model-generated probability distribution and the true label distribution in classification tasks. Additionally, it utilizes the cosine similarity loss function to assess the similarity between samples within the feature space, promoting alignment of frames from the same category and separation of frames from distinct categories. Through the combined use of these loss functions, a synergistic optimization effect enhances the feature representation capabilities of the proposed model.

4. Experimental results and analysis

4.1. Datasets and experimental environment

4.1.1. Datasets

UCF101-24 and JHMDB-21 primarily focus on sports and fitness activities. Although they include anomalous behaviors, each video in these datasets features only one individual performing actions without considering multi-person interactions. The AVA dataset can be categorized into three major types: human actions, object manipulation, and person interactions. Each frame may involve multiple actors, each possibly performing various actions. These datasets, however, are not intended for use in surveillance. This implies that they might not have certain behaviors and features, including car movement and pedestrian flow, nor realistic, everyday surveillance contexts, such as streets or malls. As such, they might not be appropriate for modeling training in surveillance environments. Although pertinent to surveillance settings, UCF-Crime and RWF-2000 only offer labels at the video

level. This limitation hinders the model's ability to learn the temporal and spatial positioning of actions and limits their practical application in real-world settings.

To facilitate comprehensive performance assessment and align with practical application scenarios, this study introduces the Public Security Abnormal Action Dataset (PSA-Dataset). Curated from 118 original videos from the UCF-Crime and RWF-2000, the PSA-Dataset notably integrates frame-level spatial annotations, ensuring comprehensive coverage in pivotal frames. It is divided into PSAD for spatio-temporal action detection and PSAR for action recognition. This dual division minimizes false negatives during model training and enhances accuracy. Each action in the dataset is endowed with two distinct labels: one designates it as an abnormal action, while the other specifies the specific category, thus evaluating the competence of the model in both abnormality detection and multi-class classification.

PSAD undergoes data preprocessing, which includes cropping, frame sampling, spatial annotations for each individual in frames, and labeling. The dataset includes annotated abnormal actions displayed in Figure 5. Within the images, green bounding boxes indicate normal actions, while red bounding boxes highlight abnormal actions.

Figure 5. Partial images in the PSAD.

As for PSAR, action tubes are constructed from the original video using PSAD labels. Each keyframe is cropped based on the ground-truth bounding box of each actor, with careful extensions to prevent the loss of action backgrounds that could impact model training. Images are resized to 224 × 224 for model data input. Cropped segments of the same actor from each keyframe are merged to form the action tube, as depicted in Figure 6. After constructing the action tube, the TV-L1 optical algorithm computes the optical images corresponding to each action tube.

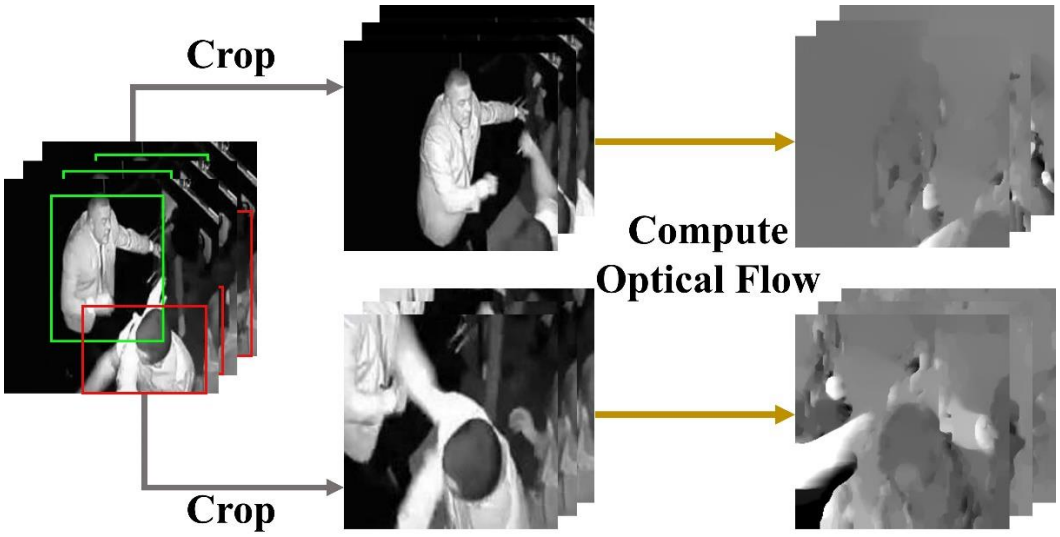

Figure 6. The production process of PSAR.

4.1.2. Experimental environment

The experiments in this study were conducted on a 64-bit Windows 10 operating system, utilizing an Intel Xeon Gold 5118 CPU running at 2.30 GHz, GPU acceleration provided by an NVIDIA Tesla V100, and 32 GB of available memory.

The performance evaluation of the model in spatio-temporal action detection employs the UCF101-24 public dataset for ablation and comparative experiments. For the action detection tasks involving the UCF101-24 dataset, the study maintains the configuration prescribed by the authors of YOWOv2 to ensure the objectivity of the experimental results. The training and testing sets are divided in an 8: 2 ratio for the PSAD. A weight decay of 0.0005 is used in conjunction with an initial learning rate of 0.0004 during training. The model is trained for a total of 50 epochs with an image batch size of 64. At the second, fourth, sixteenth, and twenty-fourth epochs, the learning rate is halved. The batch size is kept at 64 throughout the testing process.

In the context of action recognition, ablation and comparative experiments rely on the UCF101 and HMDB51 public datasets. For these datasets, this paper adheres to the official configurations. For PSAR, the study employs a RGB image batch size of 30 and an optical image batch size of 10. For RGB images, the number of channels is 3, while for optical flow images, the number of channels is 10. The initial learning rate is set at 0.001, and the model undergoes training for 40 epochs, with the learning rate halving every 2 epochs to mitigate overfitting. To mitigate overfitting, the momentum is set to 0.9, and the weight decay rate is set to 0.0005. Stochastic gradient descent is employed as the optimizer to enhance the model's robustness. To determine the optimal value of λ, this paper conducted relevant ablation experiments, which are included in the supplementary materials. The joint loss function is utilized, with a weight of 0.60 for λ and 0.40 for η. Input images are resized, center-cropped to dimensions of 224 × 224, and subject to random horizontal rotation with a 50% probability. Subsequently, images are normalized and subjected to regularization in accordance with specified parameters.

To address potential overfitting, this study employs cross-modal pretraining techniques. In the experiments, model weights pretrained on the ImageNet dataset were loaded into the baseline model to achieve thorough optimization of network parameters. For optical flow data, linear transformation is used to discretize the optical flow data into the 0–255 range, aligning it with the value domain of a single RGB channel. The weights of the first convolutional layer of the ViT are averaged and then replicated according to the channel count of the optical flow data. The input channel number of ViT's first convolutional layer is modified, and the averaged weights are loaded. Regarding regularization techniques, on top of layer normalization and Dropout in ViT, an additional Dropout layer is added after the final layer normalization to further reduce the impact of overfitting, with a Dropout rate set at 0.5. For data augmentation techniques, scale jittering, corner cropping, and random horizontal flipping are included. These techniques help enhance the model's ability to recognize data from different angles and positions. During testing, both RGB and optical images are processed with a batch size set to 1. Each video is evenly divided into 25 segments, from which a single frame is randomly selected to produce 25 frames. All images are scaled, center-cropped to 224 × 224, normalized, and standardized, yielding 25 images of 224 × 224 each. The primary metric for assessing the performance of the action recognition algorithms in this paper is Accuracy.

4.2. Experimental results of spatio-temporal action detection

This paper rigorously validates the performance of the YOWOv2E model in the domain of general action detection tasks, utilizing comparisons with contemporary leading action detection models on the UCF101-24 dataset, as meticulously outlined in Table 1. YOWOv2E emerges as a standout performer in terms of detection accuracy on this widely recognized public dataset while maintaining exceptional computational efficiency. Notably, in contrast to its counterparts, including ACT [26], MOC [28], and YOWO [9], YOWOv2E showcases a distinctive advantage by operating with significantly lower floating-point operations, thus enabling expedited inference. This attribute greatly enhances its applicability in real-world scenarios and substantially multi-gates computational complexity, addressing a pressing concern in the field of action detection. Furthermore, YOWOv2E distinguishes itself by surpassing its predecessor, YOWOv2, showcasing substantial enhancements in general action detection tasks. This includes improvements in target localization and classification while preserving the original computational efficiency of the proposed model. The YOWOv2E model strikes an admirable equilibrium between performance and computational costs, endowing it with a distinct practical advantage, particularly in resource-constrained scenarios encountered in real-world contexts such as public safety surveillance.

Additionally, the YOWOv2E approach demonstrates remarkable robustness and versatility when contrasted with other leading models. Its consistent, stable performance extends across diverse scenarios and a spectrum of abnormal actions, making it an attractive choice for a wide range of applications. In essence, Table 1 effectively substantiates the superior performance of the YOWOv2E model on the UCF101-24 dataset. Not only does it surpass its competitors in terms of accuracy, but it also unambiguously exhibits clear advantages in computational efficiency, thereby forming a solid foundation for the extensive deployment of this approach in various public safety monitoring applications.

Table 1. Performance comparison of different methods on UCF101-24.

Method	Frame-mAP/%	GFLOPs ($\times 10^9$)
T-CNN [25]	41.4	-
ACT [26]	67.1	256.2
STEP [27]	75	125.7
MOC [28]	77.8	> 93.2
YOWO [9]	77.8	39.3
YOWOv2 [10]	77.14	1.28
YOWOv2E (Ours)	**78.52**	**1.28**

Abnormal actions often present distinctive characteristics which significantly intensify the challenge of accurately distinguishing them from normal ones. In response to this intricate classification challenge, this paper introduces the YOWOv2E model, an extension of the YOWOv2. This extended model introduces innovative components, namely the ShuffleNetv2E feature extraction network and a joint loss function. By capitalizing on advanced techniques like the ECA mechanism, Focal Loss, and label smoothing, the model greatly amplifies its capabilities in feature extraction and classification, with a particular focus on abnormal actions.

Table 2 provides an insightful view of the experimental results related to binary classification, precisely discriminating between normal and abnormal actions using both YOWOv2 and YOWOv2E. A meticulous review of Table 2 underscores the exceptional proficiency of YOWOv2E in locating and categorizing actions across both abnormal and normal action detection scenarios. Notably, YOWOv2E excels with a remarkable 5.77% improvement in Frame-mAP for abnormal action recognition compared to its predecessor, YOWOv2. This outcome underscores heightened capacity of YOWOv2E to extract features associated with abnormal actions and refine the boundaries of classification between abnormal and normal actions, ultimately enhancing the precision of abnormal action detection.

Table 2. Comparison of binary classification results on PSAD.

Method	Abnormal/%	Normal/%	Total/%
YOWOv2	62.10	78.59	70.34
YOWOv2E (ShuffleNetv2E)	63.99	75.34	69.67
YOWOv2E (Improved Loss)	65.19	77.81	71.50
YOWOv2E	**67.87**	**81.86**	**74.87**

Furthermore, Table 3 provides an in-depth presentation of multi-class classification experiments executed on the PSAD dataset. YOWOv2E shines, delivering a substantial 2.73% enhancement in Frame-mAP when compared to YOWOv2. This success can be attributed to the integration of the ShuffleNetv2E module, combined with the ECA mechanism. This strategic amalgamation empowers the model to focus intently on critical information within video frames, dynamically adjusting the weights assigned to each input. The relevance of this feature within the realm of video analysis becomes apparent, as it accommodates the diverse and varied nature of actional information embedded in different frames.

Moreover, the joint loss function plays a pivotal role in bolstering the classification capabilities of the model across a spectrum of actions. This function effectively expands the boundaries of action classification, instills greater confidence in action detection, and systematically enhances the capability to address class imbalances.

Table 3. Comparison of multiple classification results on PSAD.

Method	Frame-mAP/%
YOWOv2	49.45
YOWOv2E (ShuffleNetv2E)	50.96
YOWOv2E (Improved Loss)	51.50
YOWOv2E	**52.18**

The paper further delves into a comparative analysis of the proposed models, scrutinizing their performance in detecting specific categories. Visual results of abnormal action detection by YOWOv2 and YOWOv2E are vividly presented in Figure 7. Action detection grapples with a unique challenge – the blurred boundary that distinguishes abnormal from normal behaviors. This intrinsic ambiguity can lead to classification errors, including both missed detections and false alarms, as illustrated in Figure 7(a),(d). YOWOv2, for instance, misclassifies actions such as theft and fighting as normal activities like walking and standing. In stark contrast, YOWOv2E, harnessing the potential of ShuffleNetv2E and the joint loss function, proficiently identifies abnormal actions while imbuing a heightened level of confidence in its predictions.

The challenge of distinguishing abnormal actions from normal ones, often complicated by blurred boundaries, is vividly depicted in Figure 7(a). Notably, YOWOv2 incorrectly categorizes a fall as a fight action, highlighting the classification shortcomings. In stark contrast, YOWOv2E excels by precisely recognizing abnormal actions, effectively addressing the intricacies of boundary definition. YOWOv2E takes a step further by intensifying feature extraction across the entire image, prioritizing image details, and significantly reducing the chances of overlooking actions, as showcased in Figure 7(e). This enhancement is attributed to the incorporation of the ECA mechanism and meticulous loss function optimization, collectively elevating the understanding of complex scenes. Consequently, the model adeptly captures spatio-temporal features associated with abnormal actions in surveillance videos. This progress not only facilitates early anomaly detection but also empowers the extraction of crucial evidence, presenting substantial potential for enhancing law enforcement efficiency and bolstering public safety efforts.

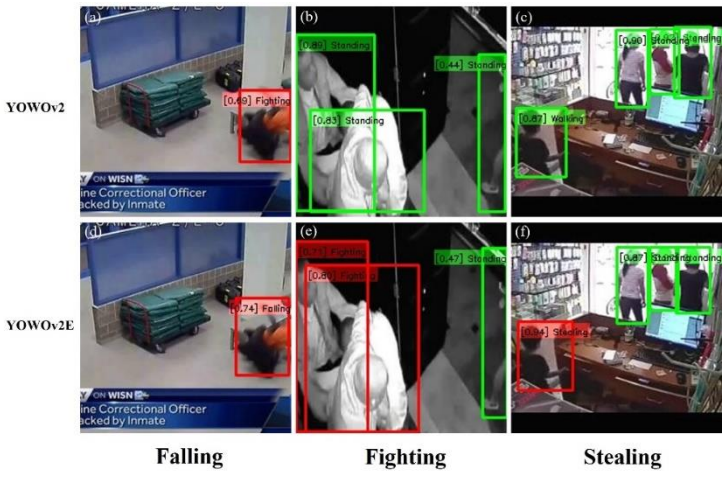

Figure 7. Display of the detection results. (a)–(c) detection results of YOWOv2; (d)–(f) detection results of YOWOv2E.

4.3. Comparative experiments of different models

The difference in Loss for various baselines when trained on UCF101 is seen in Figure 8. The model converges more quickly when the baseline is swapped out for the ViT, and the Loss post-convergence also significantly decreases. The model's convergence speed is further increased and the loss is further decreased by modifying the loss function and adding SimAM.

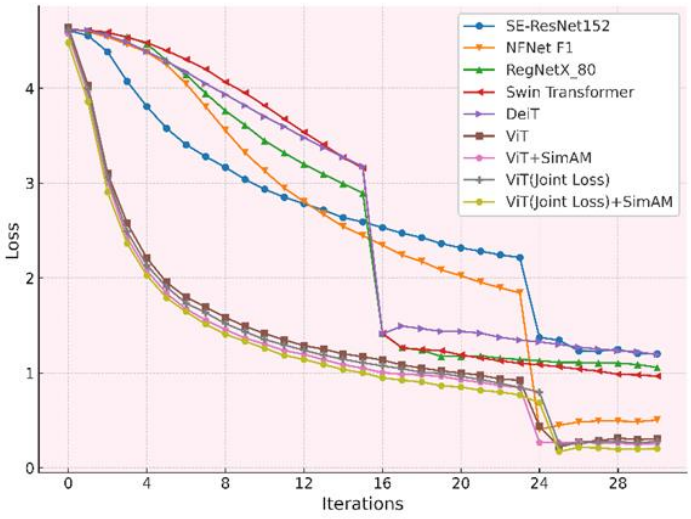

Figure 8. Comparison of loss of different baselines on UCF101.

For action recognition tasks, important factors include the actor's appendages and related objects, whereas irrelevant factors include the action's background. The more important factors and the fewer irrelevant factors the model learns, the better it performs. As depicted in Figure 9, embedding SimAM increases the model's attention to the appendages of people and objects related to actions and reduces the interference of action backgrounds, thereby enhancing the model's ability to represent features.

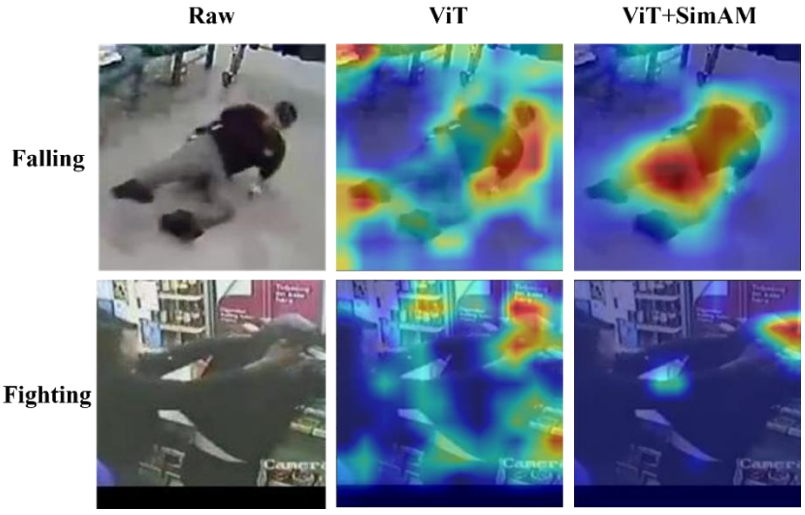

Figure 9. Visual feature heat maps.

The ViT introduces a self-attention mechanism that effectively identifies long-range dependencies among image regions, effectively bypassing the constraints associated with the local receptive fields in CNN. By integrating transfer learning techniques and capitalizing on pretrained weights designed for image classification tasks, ViT mitigates the impact of the lack of inductive bias, yielding substantial performance improvements in action recognition tasks. The comparative analysis presented in Table 4 underscores the superior performance of ViT when contrasted with two CNN-based neural networks, SE-ResNet152 [29] and NFNet-F1 [30]. The latter networks excel primarily in image classification tasks but exhibit limitations stemming from their inductive biases, such as translation invariance and local sensitivity, which hinder their ability to comprehensively capture global image information and evaluate feature interdependencies. As a consequence, their understanding of the overall image context remains incomplete.

In contrast, the self-attention mechanism of ViT empowers the network to concentrate on relationships among any image blocks, allowing it to effectively utilize contextual information in modeling global dependencies within images. This pivotal feature compensates for the limitations inherent in convolutional architectures.

Table 4. Comparison of results of different baselines on UCF101.

Baseline	Spatial/%	Temporal/%	Two-Stream/%	Param ($\times 10^6$)	GFLOPs ($\times 10^9$)
SE-ResNet152	83.06	72.55	89.65	65	148.06
NFNet-F1	85.97	83.22	93.19	129.87	226.8
ViT	87.28	83.37	94.59	85.7	228.36
ViT + SimAM	88.58	84.97	95.35	85.7	228.36
ViT (Joint Loss)	88.79	84.21	95.1	85.7	228.36
ViT (Joint Loss) + SimAM	**89.67**	**85.33**	**95.63**	**85.7**	228.36

4.4. Comparative experiments of temporal attention and number of segments

In order to comprehensively extract temporal features, this paper has introduced a temporal self-attention module into the proposed model, and this work has undertaken a series of ablation experiments to evaluate its optimal placement. As depicted in Table 5, when the temporal self-attention module is positioned at the beginning of the ViT, it prematurely amalgamates temporal and visual information, leading to increased complexity in the grasp of interframe relationships about the model. Conversely, when the temporal self-attention module is located at the end of ViT, it ensures that the model first acquires sufficient abstract visual semantic information before incorporating temporal data. This strategic positioning empowers the temporal self-attention mechanism to enhance the model's understanding of relationships among distinct time frames within the input sequence. Consequently, this arrangement facilitates a more accurate capture of crucial information within the sequence. Given that action recognition fundamentally relies on a profound comprehension of inter-frame relationships, the inclusion of temporal self-attention at the end of ViT consistently outperforms its alternative placements.

Moreover, this study has conducted experiments employing 3, 6 and 9 segments to investigate the impact of segment count on model accuracy. As indicated in Table 5, video segmentation surpasses the performance of non-segmented data, emphasizing that segmenting videos augments the efficiency of the model in handling prolonged sequential data. However, it is imperative to note that an excessive number of segments can impede the effective extraction of temporal features and result in heightened

computational demands.

Table 5. Comparison of RGB results of different segments and whether to add temporal attention on UCF101.

Num of segments	Head/%	End/%	None/%	GFLOPs ($\times 10^9$)
1Frame (RGB)	—	—	85.03	17.58
3Frames (RGB)	88.23	**90.12**	89.67	52.74 (+ 4.16)
6Frames (RGB)	85.72	88.71	86.55	105.42 (+ 8.41)
9Frames (RGB)	84.85	87.39	86.36	158.13 (+ 12.63)
3Frames (Optical)	85.49	**86.27**	85.33	175.62 (+ 41.69)
3Frames (RGB + Optical)	95.46	**96.12**	95.63	228.36 (+ 45.85)

4.5. Results comparison on public datasets

Table 6 provides a comprehensive overview of the experimental results at each stage of our study. As this paper advanced through the precise detection stage, we consistently observed a noteworthy increase in accuracy, unequivocally affirming the effectiveness of the enhancements made.

Table 6. Results of the ablation validation experiment on UCF101.

Model	Accuracy/%
ViT (RGB)	86.35
ViT (RGB + Optical)	94.23
ViT + Decision Fusion (0.6)	94.59
ViT + SimAM	95.35
ViT + Joint Loss	95.63
ViT + Temporal Attention	**96.12**

Table 7. Comparison of accuracy of different models on UCF101 and HMDB51.

Model	UCF101/%	HMDB51/%
iDT + FV	85.93	57.2
iDT + HSV	87.92	61.11
Two-Stream	88.02	59.4
ResNext-101 + SE + SA (16f) [31]	92.5	68.0
TSN	92.64	65.74
HAR-depth [32]	93.0	69.7
spatio-temporal STFT [33]	94.7	71.5
HME-Net [34]	94.8	72.2
BifurcatedNet [35]	94.9	72.1
ViTSN	**96.12**	**73.5**

To ascertain the superiority of the proposed method, this work conducted a rigorous comparative analysis of its recognition performance against other widely adopted methodologies. As depicted in Table 6, the proposed model outperforms the prevailing algorithms currently in use. When compared

to TSN, the proposed approach demonstrates substantial improvements in accuracy, achieving a noteworthy enhancement of 3.48% on UCF101 and 7.76% on HMDB51. This underscores the ability of the method to significantly elevate the accuracy of models in the realm of action recognition, highlighting the numerous advantages inherent in the approach presented in this paper.

4.6. Results comparison on PSAR of ViTSN

ViTSN undergoes a rigorous comparative evaluation against the baseline TSN framework on the PSAR dataset to assess its generalization capabilities. As detailed in Table 8, the results affirm the suitability of the proposed model for real-world applications, particularly in the context of public security. Compared to TSN, the method proposed in this paper first embeds non-parametric attention modules and temporal attention modules at the beginning and end of the network. In order to evaluate feature weights, it uses a more thorough and effective attention mechanism, which improves the model's capacity to handle background noise and lengthy temporal sequence data. Second, it fully captures the spatial relationships between image block features by utilizing a spatio-temporal feature extraction module based on pretrained ViT, which produces more representative image feature vectors. To further enhance the model's classification capabilities, the training loss function has been adjusted. It highlights the aptitude of the model for finegrained differentiation of abnormal actions and the extraction of specific video evidence from extensive monitoring data. Consequently, this approach shows some benefits by increasing accuracy in behavior recognition tests.

Table 8. Comparison of accuracy of different models on PSAR.

Model	RGB/%	Flow/%	Two Stream/%
TSN	32.04	35.4	42.99
ViTSN	**51.35**	**40.81**	**51.79**

Additionally, in Figure 10, this work presents the confusion matrices for both models, encompassing the recognition of seven distinct action classes within the test dataset. A noteworthy distinction emerges when comparing the proposed model to the baseline. The method excels in the recognition of abnormal actions and exhibits superior performance in identifying abnormal action classes, significantly enhancing its applicability to the practical demands of public security work.

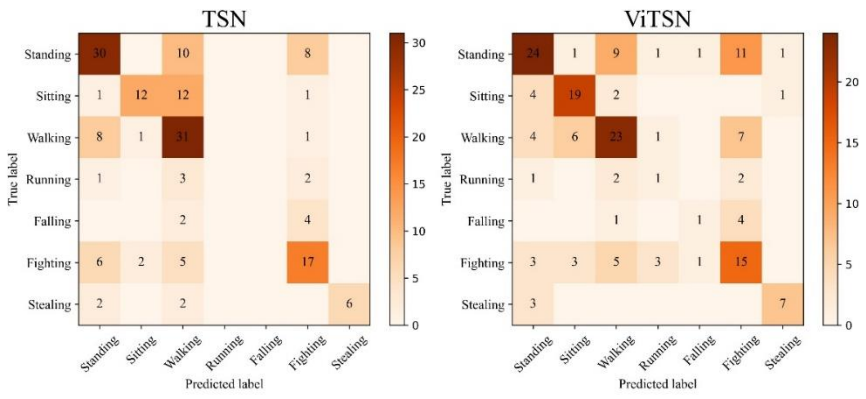

Figure 10. Comparison of confusion matrices of different models on PSAR.

5. Conclusions

This paper presents YOWOv2E, an advanced lightweight model for video abnormal action detection tailored for forensic science. YOWOv2E integrates ShuffleNetv2E for feature extraction, is augmented with the ECA mechanism, and employs a joint loss function to significantly enhance the accuracy of detecting abnormal actions in complex surveillance scenarios. Additionally, this paper introduces a spatio-temporal two-stream network model based on ViT, incorporating the SimAM attention mechanism to enhance the model's resilience to interference. To handle long sequential data, segmental sampling strategies are implemented, and decision layer fusion is employed to improve accuracy. The effectiveness of the proposed model is further validated on the PSA-Dataset, a novel surveillance abnormal action dataset developed in this study, highlighting its robust generalization performance. This research aligns with the practical requirements of public security work and offers valuable support to agencies involved in electronic evidence investigations.

However, the sample size for some categories in this dataset is still insufficient because publicly available surveillance footage of abnormal actions is scarce, which may have hindered the model's detection capabilities. Future research could concentrate on a number of areas, including expanding the dataset's sample size and scene variety, investigating techniques based on weak supervision or self-supervision to lessen the labor-intensive nature of manual labeling, and creating more effective and lightweight model structures for simpler deployment. All of these directions merit additional research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, China No. 2023JKF01ZK05.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. K. Q. Huang, X. T. Chen, Y. F. Kang, T. N. Tan, Intelligent visual surveillance, *Chin. J. Comput.*, **38** (2015), 1093–1118. http://dx.doi.org/10.11897/SP.J.1016.2015.01093
2. E. Selvi, M. Adimoolam, G. Karthi, K. Thinakaran, N. M. Balamurugan, R. Kannadasan, et al., Suspicious actions detection system using enhanced CNN and surveillance video, *Electronics*, **11** (2022), 4210. https://doi.org/10.3390/electronics11244210
3. M. Jain, J. V. Gemert, H. Jégou, P. Bouthemy, C. G. M. Snoek, Action localization with tubelets from motion, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014), 740–747. http://doi.ieeecomputersociety.org/10.1109/CVPR.2014.100

4. K. Soomro, H. Idrees, M. Shah, Action localization in videos through context walk, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 3280–3288. https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.375
5. G. Gkioxari, J. Malik, Finding action tubes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 759–768. https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298676
6. G. Singh, S. Saha, M. Sapienza, P. Torr, F. Cuzzolin, Online real-time multiple spatiotemporal action localisation and prediction, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 3637–3646. https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.393
7. R. Hou, C. Chen, M. Shah, Tube convolutional neural network (T-CNN) for action detection in videos, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 5822–5831. https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.620
8. C. H. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Q. Li, et al., AVA: A video dataset of spatio-temporally localized atomic visual actions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 6047–6056. https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00633
9. O. Köpüklü, X. Y. Wei, G. Rigoll, You only watch once: A unified cnn architecture for real-time spatiotemporal action localization, preprint, arXiv:1911.06644. https://doi.org/10.48550/arXiv.1911.06644
10. J. H. Yang, K. Dai, YOWOv2: A stronger yet efficient multi-level detection framework for real-time spatio-temporal action detection, preprint, arXiv:2302.06848. https://doi.org/10.48550/arXiv.2302.06848
11. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. F. Li, Large-scale video classification with convolutional neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014), 1725–1732. https://doi.ieeecomputersociety.org/10.1109/CVPR.2014.223
12. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, *Adv. Neural Inf. Process. Syst.*, **27** (2014). https://doi.org/10.48550/arXiv.1406.2199
13. L. M. Wang, Y. J. Xiong, Z. Wang, Y. Qiao, D. H. Lin, X. O. Tang, et al., Temporal segment networks for action recognition in videos, *IEEE Trans. Pattern Anal. Mach. Intell.*, (2016), 20–36. https://doi.org/10.1007/978-3-319-46484-8_2
14. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 6836–6846. https://doi.org/10.1109/ICCV48922.2021.00676
15. H. Q. Fan, B. Xiong, K. Mangalam, Y. H. Li, Z. C. Yan, J. Malik, et al., Multiscale vision transformers, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 6824–6835. https://doi.org/10.1109/ICCV48922.2021.00675
16. G. Bertasius, H. Wang, H. Torresani, Is space-time attention all you need for video understanding?, in *Proceedings of the 38th International Conference on Machine Learning*, (2021), 813–824. https://arxiv.org/abs/2102.05095
17. S. Yan, X. H. Xiong, A. Arnab, Z. C. Lu, M. Zhang, C. Sun, et al., Multiview transformers for video recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 3333–3343. https://doi.org/10.1109/CVPR52688.2022.00333

18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, **30** (2017), 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349
19. L. Yang, R. Y Zhang, L. Li, X. Xie, Simam: A simple, parameter-free attention module for convolutional neural networks, in *International Conference on Machine Learning, PMLR*, (2021), 11863–11874. https://proceedings.mlr.press/v139/yang21o
20. C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
21. N. N. Ma, X. Y. Zhang, H. T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 116–131. https://doi.org/10.1007/978-3-030-01264-9_8
22. Q. L. Wang, B. G. Wu, P. F. Zhu, P. H. Li, W. M. Zuo, Q. H. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155
23. T. Y. Lin, P. Goyal, R. Girshick, K. M. He, P. Dollár, Focal loss for dense object detection, in *IEEE International Conference on Computer Vision (ICCV)*, (2020), 2999–3007. https://doi.org/10.1109/ICCV.2017.324
24. R. Müller, S. Kornblith, G. Hinton, When does label smoothing help?, *Adv. Neural Inf. Process. Syst.*, **32** (2019), 4694–4703 https://dl.acm.org/doi/10.5555/3454287.3454709
25. R. Hou, C. Chen, M. Shah, Tube convolutional neural network (T-CNN) for action detection in videos, in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 5822–5831. https://doi.org/10.1109/ICCV.2017.620
26. V. Kalogeiton, P. Weinzaepfel, V. Ferrari, C. Schmid, Action tubelet detector for spatio-temporal action localization, in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 4405–4413. https://doi.org/10.1109/ICCV.2017.472
27. X. T. Yang, X. D. Yang, M. Y. Liu, F. Y. Xiao, L. S. Davis, J. Kautz, STEP: Spatio-temporal progressive learning for video action detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 264–272. https://doi.org/10.1109/CVPR.2019.00035
28. Y. X. Li, Z. X. Wang, L. M. Wang, G. S. Wu, Actions as moving points, in *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, (2020), 68–84. https://doi.org/10.1007/978-3-030-58517-4_5
29. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
30. A. Brock, S. De, S. L. Smith, K. Simonyan, High-performance large-scale image recognition without normalization, in *International Conference on Machine Learning, PMLR*, **139** (2021), 1059–1071. https://doi.org/10.48550/arXiv.2102.06171
31. F. Anvarov, D. H. Kim, B. C. Song, Action Recognition Using Deep 3D CNNs with sequential feature aggregation and attention, *Electronics*, **9** (2020), 147. https://doi.org/10.3390/electronics9010147

32. S. P. Sahoo, S. Ari, K. Mahapatra, S. P. Mohanty, HAR-depth: a novel framework for human action recognition using sequential learning and depth estimated history images, *IEEE Trans. Emerging Top. Comput. Intell.*, **5** (2020), 813–825. https://doi.org/10.1109/tetci.2020.3014367
33. S. Kumawat, M. Verma, Y. Nakashima, S. Raman, Depthwise spatio-temporal STFT convolutional neural networks for human action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 4839–4851. https://doi.org/10.1109/TPAMI.2021.3076522
34. B. Wang, X. H. Wang, S. W. Ren, W. J. Wang, Y. T. Shi, Hierarchical motion excitation network for few-shot video recognition, *Electronics*, **12** (2023), 1090. https://doi.org/10.3390/electronics12051090
35. J. X. Zhang, H. F. Hu, Z. Liu, Appearance-and-dynamic learning with bifurcated convolution neural network for action recognition, *IEEE Trans. Circuits Syst. Video Technol.*, **31** (2020), 1593–1606. https://doi.org/10.1109/TCSVT.2020.3006223

AIMS Press