IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 August 2023, accepted 11 September 2023, date of publication 18 September 2023,
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316724

RESEARCH ARTICLE

Spatio-Temporal Attention Fusion SlowFast for Interrogation Violation Recognition

HAILUN WANG[1,2], **BIN DONG**[1,2], **QIRUI ZHU**[1,3], **ZHIQIANG CHEN**[1], **AND YI CHEN**[1,4], **(Senior Member, IEEE)**

[1]College of Electrical and Information Engineering, Quzhou University, Quzhou, Zhejiang 324000, China
[2]School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310000, China
[3]School of Computing, Hangzhou Dianzi University, Hangzhou, Zhejiang 310000, China
[4]School of Engineering, Newcastle University, NE1 7RU Newcastle upon Tyne, U.K.

Corresponding author: Zhiqiang Chen (czq@qzc.edu.cn)

This work was supported in part by the Natural Science Foundation of Zhejiang Province under Grant LZY22E050003; in part by the Basic Public Welfare Research Plan Project of Zhejiang Province funded by the Science Technology Department of Zhejiang Province, China, under Grant LGG22F010011; and in part by the Startup Research Found Plan Project funded by Quzhou University, China, under Grant BSYJ202107.

ABSTRACT The use of video surveillance to monitor interrogation behavior can effectively maintain judicial civility in the context of law enforcement cases. However, analyzing and reviewing law enforcement videos can be a time-consuming and resource-intensive process, particularly in the manual identification of interrogation violations. This work is dedicated to the development of an intelligent recognition system for interrogation violations by using a spatio-temporal attention fusion SlowFast Network. To address the issue of feature information underutilization in the slow path of the traditional SlowFast, a slow-to-fast path is incorporated into the original SlowFast to enhance learning. The model fuses the attention of spatial and temporal channels, replacing the traditional convolution module with this new approach. The proposed model was evaluated using the publicly available UCF101 action recognition dataset, resulting in a 1.52% improvement in Top-1 recognition accuracy compared to the traditional SlowFast. Based on two custom interrogation misconduct datasets, the proposed model was evaluated, achieving a recognition rate of 99.16% for interrogation misconduct. This demonstrates its effectiveness in identifying misconduct behaviors inside interrogation rooms. Compared to some advanced behavior recognition models, the proposed model demonstrates strong competitiveness in identifying misconduct during interrogations.

INDEX TERMS Attention fusion, enhance learning, law enforcement, SlowFast, violation recognition.

I. INTRODUCTION

The prohibition of using torture to extract confessions is a fundamental aspect of civilized interrogation practices, and many countries have implemented laws and regulations that strictly forbid its use. For instance, China's Criminal Law and Criminal Procedure Law, enacted in 2013, explicitly prohibit the use of torture to extract confessions. Despite these legal measures, the use of torture for this purpose remains a significant problem, presenting a challenge for judicial oversight agencies worldwide in curbing its occurrence. This is crucial in upholding judicial justice and safeguarding the legitimate rights and interests of criminal suspects. In addition to the establishment of relevant laws and regulations, the use of technological means to prevent and detect improper behavior by interrogation officers in real-time has become a powerful measure. Therefore, audio and video surveillance equipment has become standard in interrogation rooms.

Some scholars have conducted relevant research on the recognition of improper interrogation behavior by law enforcement personnel. Wang [1] constructed a 3D model of the human body in the interrogation scene and utilized 3D spatial scene and 2D video feature constraints for the recognition of misconduct such as beating, hanging, and

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

binding in the interrogation room. Li et al. [2] enhanced the network's ability to recognize non-compliant behavior by adding dense connection blocks to the traditional network to aid in the learning of temporal features. Tan et al. [3] combined 3D convolution and LSTM, using LSTM to model the short-term features extracted by 3D convolution along the time axis, assisting in the recognition of violent behavior in public security monitoring videos. However, most existing algorithms for recognizing improper interrogation behavior preselect human behavioral features, which have drawbacks such as scene dependency and algorithm dependency. From a video perspective, there is a lack of real-time modeling capability for temporal dynamics, which severely hinders the promotion and application of intelligent surveillance.

This research comes from the real needs of the public security system in China, focusing on the detection of assault, long squat, push-up, abnormal running, abnormal jumping, assault with stick and normal behavior. The corresponding categories contain violent behaviors, long time behaviors and normal behaviors, with obvious differences in timing feature information, which tests the performance of the model more. In order to improve the efficiency of the detection, relevant improvements have been made to the network, which have made significant contributions. Firstly, a new mechanism for feature extraction from different spatial locations in the Slow path of the SlowFast network has been developed, as well as a mechanism for fusing feature information with the Fast path, in order to better handle spatial information. Secondly, a weight adjustment adaptive mechanism has been proposed for each Fast path to Slow path information to improve the model's attention to different channels. Finally, two customized interrogation behavior datasets have been constructed, containing normal behaviors and various types of violations, and an in-depth performance evaluation of the proposed model has been conducted based on these datasets.

The paper is organized as follows. Section II introduces the development and related research in the field of violation behavior recognition, including traditional handcrafted methods and deep learning methods. Section III describes the specific details of the proposed new model architecture. Section IV primarily presents the experiments conducted on a publicly available dataset. Section V focuses on the collection of interrogation behavior dataset and performance evaluation. Finally, Section VI summarizes the contributions of the paper and outlines the future work objectives.

II. RELATED WORK

The core of interrogation violation recognition is the recognition of human behavior. With the advancement of deep learning, computer vision has made significant progress and has been applied to various fields, including human behavior recognition, which is a critical application area. Human behavior recognition involves understanding the image content, which is more challenging than recognizing or detecting objects in an image due to the diverse and complex human poses and factors such as occlusion and background clutter. Various human behavior recognition methods have been developed as research hotspots. Based on the feature extraction method, these methods can be categorized into traditional manual-based methods [4], [5], [6] and deep learning-based behavioral feature extraction methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18].

A. TRADITONAL METHODS

The traditional method of manually extracting features for human behavior recognition can be classified into two types: key point and motion trajectory extraction. Local key point model features are often described using spatiotemporal interest points, which are key points that change significantly in both space and time during motion. Schuldt et al. [4] proposed a descriptor that uses high-order representation features to construct a video representation based on local spatiotemporal features, and combined this representation with an SVM classification scheme to recognize motion patterns. Lowe [5] introduced the classical SIFT (Scale Invariant Feature Transform) descriptor based on the scale-invariant property, which can be used to reliably match between different views of an object or scene. Scovanne et al. [6] combined the spatiotemporal relationship properties of video images and introduced the SIFT operator for video-based human behavior analysis. This new descriptor can better represent the 3D properties of video data in action recognition applications. However, the traditional manual feature extraction method has limitations, such as being restricted to specific needs due to small data in the initial sample database, simple scenes, and single actions. Moreover, manual feature extraction is time-consuming and inefficient, requiring significant manpower. External factors can easily disturb the effect of manual feature extraction, making it unstable, and the recognition accuracy has much room for improvement.

B. DEEP LEARNING METHODS

Since the emergence of AlexNet [7], deep learning has played a crucial role in the field of computer vision. Convolutional Neural Networks (CNNs) have been employed for feature extraction on videos. However, recognizing each frame of a video with a CNN alone is not sufficient, as videos have an additional temporal dimension that influences the semantic content. Researchers have explored various methods to address this challenge. Karpathy et al. [8] studied various methods for extending CNN connections in the temporal domain to leverage local spatiotemporal information. They attempted four different methods for fusing cross-temporal information, but the results were not satisfactory. Simonyan and Zisserman [9] proposed using two channels with different structures to process spatial and temporal information separately, and fused the information at the end of the process, yielding competitive results with manual feature extraction. Regarding information fusion in dual-stream networks, Feichtenhofer et al. [10] investigated fusion strategies for network architectures and studied various methods for spatial and temporal fusion of ConvNets in order to best utilize spatiotemporal information. They found that early fusion of features continued to yield better results. Tran et al. [11] proposed a 3D convolutional neural network, which extended the 2D convolutional kernel to 3D by increasing the temporal dimension, to handle spatiotemporal information feature extraction. This approach outperformed traditional 2D convolutional networks. Feichtenhofer et al. [12] developed the SlowFast model, which used both slow and fast network branches to achieve efficient video recognition, significantly improving the accuracy and speed of video understanding tasks. Attention mechanisms have also been applied in the vision domain. Arnab et al. [13] used pure Transformer modules to address classification problems in the video domain. However, compared to CNN, the model has fewer inductive biases and requires pre-training to achieve good results. Ge et al. [14] proposed a convolutional LSTM action recognition algorithm based on attention mechanism, aiming to enhance the accuracy of action recognition by effectively extracting salient regions of actions in videos. Bertasius et al. [15] extended the vision transformer from the image domain to the video domain by directly learning spatiotemporal features from a sequence of frame-level patches. They adapted the standard Transformer architecture to videos and proposed the Timesformer architecture based on spatiotemporal attention mechanisms. Compared to 3D convolutional networks, this model trains faster and achieves higher testing efficiency. Sharma et al. [16] proposed a video action recognition model based on soft attention, which focuses on identifying important elements in video frames based on the ongoing actions. Patrick et al. [17] proposed trajectory attention, which aggregates information along the implicitly determined motion path, to better capture the temporal information contained in videos and effectively assist video understanding. Xiang et al. [18] proposed a simple and efficient temporal self-attention transformer, which implements a spatiotemporal self-attention mechanism without increasing computation or the number of parameters compared to 2D transformer networks.

Interrogation violations often involve violent behavior, making it an important area of research in behavior recognition. However, detecting violent behavior poses a significant challenge due to its volatile, fast, and difficult-to-capture nature. Compounding this challenge is the lack of a normative assessment guide to aid in the identification of violent behavior. To address this issue, researchers have explored different approaches. Datta et al. [19] used the motion trajectories and limb orientations of people in a scene to detect violence. They used the motion trajectory and direction information of a person's four limbs to detect human violence in videos, such as punching, kicking, and using objects to strike. Lam et al. [20] evaluated the use of multiple features and their combinations in a violence scene detection system, providing an empirical basis for selecting capable feature sets to deal with heterogeneous content in movies that include violence scenes. Hassner et al. [21] developed a violence flow descriptor (ViF) based on the size of flow vectors, which were then classified as violent or non-violent using a support vector machine (SVM) in crowd scenes. Ding et al. [22] proposed a 9-layer 3D-CNN for violence video detection and achieved a score of 91% on a hockey game dataset. However, their work used 3D convolution but employed 2D pooling, resulting in the loss of temporal information in the input signal. Dong et al. [23] developed a multi-stream convolutional neural network framework that processes RGB images from the spatial network and optical and acceleration stream images from the temporal network separately using convolutional neural networks. The classification results are then obtained through fusion at the end of the model. Zhou et al. [24] proposed a new input modality, image acceleration field, to better extract motion attributes. First, each video is constructed as an RGB image. Second, the optical flow field is computed using consecutive frames, and the acceleration field is obtained based on the optical flow field. Third, FightNet is trained using three input modalities, namely RGB images for the spatial network, optical flow images for the temporal network, and acceleration images. By fusing the results of different inputs, violence interactions are detected by determining whether the video tells a violent event.

In summary, attention mechanisms have been widely applied in the field of behavior recognition, and their performance has been improved to some extent when combined with most existing models. Regarding the detection of interrogation misconduct, most of these behaviors have violent characteristics, and the behavior expressions may be more implicit, with less data available, making their detection much more difficult than normal behavior detection. The current focus of violence behavior detection is on how to combine spatiotemporal features for behavior detection, but there still exist problems of low detection accuracy and excessive computational complexity. Based on these ideas and issues, research on detecting misconduct in interrogation rooms can be carried out.

III. THE PROPOSED SPATIO-TEMPORAL ATTENTION FUSION SLOWFAST

In this section, a more detailed description of the proposed Spatio-Temporal Attention Fusion SlowFast (STAF-SlowFast) is provided. The main difference between images and videos lies in the temporal dimension. The traditional SlowFast model is a deep neural network that distinguishes between input paths with different sampling rates, as proposed by Feichtenhofer et al. [12]. The model is divided into slow and fast paths based on the number of input frames. The slow path mainly uses deep 2D convolution and non-degenerate temporal convolution to process spatial information, while the fast path uses shallow 3D convolution to extract short-term features in time. However, the traditional SlowFast model's one-way connection only considers the fusion of feature information from the fast path into the slow path, and it does not fully utilize the feature information

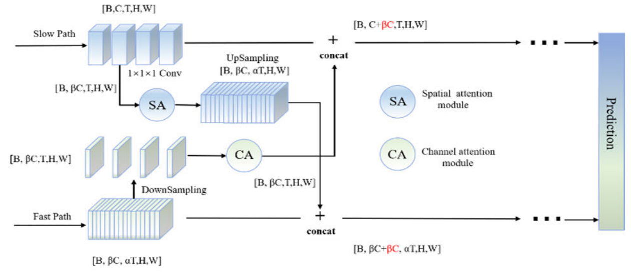

FIGURE 1. Overview of the proposed Spatio-Temporal Attention Fusion SlowFast.

in the slow path. To address this issue, a spatio-temporal attention fusion module replaces the traditional convolutional module by adding a slow-to-fast path to the original network, facilitating learning. The spatio-temporal attention fusion module comprises channel attention and spatial attention, where channel attention is the information exchange component from the fast path to the slow path, and spatial attention is the information exchange component from the slow path to the fast path. Both extract video features and fuse them into the corresponding paths. The channel attention mechanism uses adaptive weight coefficients to capture essential information from the feature maps and discard irrelevant features, which helps to improve the network's performance. The spatial attention mechanism is used to emphasize the regions of interest and reduce interference from irrelevant regions. By fusing spatial and temporal information, the STAF-SlowFast network achieves more efficient and accurate behavior recognition than the traditional SlowFast model and some existing behavior recognition networks on the public medium-sized video dataset, as well as the interrogation violation dataset.

Fig. 1 depicts the overall architecture of the proposed Spatio-Temporal Attention Fusion SlowFast (STAF-SlowFast) network. The feature map of the slow pathway is represented as $[B, C, T, H, W]$, where B is the batch size, C is the number of channels, T is the temporal duration, and H and W are the spatial height and width, respectively. Similarly, the feature map of the fast pathway is denoted as [B, βC, αT, H, W], where α and β denote the frame rate ratio and channel ratio, respectively, as defined in the original SlowFast paper by Feichtenhofer et al. [12]. The STAF module is the key contribution of the proposed network, replacing the traditional convolutional module. The next section will provide a detailed description of the STAF module.

A. SPATIAL ATTENTION

In the proposed model, Spatial Attention (SA) aims to extract features from various spatial locations in the slow path using weighted processing and fuse them with the fast path in the channel dimension, thereby enabling the network to handle spatial information more effectively and improve its performance.

The fusion method employed in this study involves reducing the channel dimension through a $1 \times 1 \times 1$ convolution operation with a kernel size, resulting in a reduction of channels to βC. Following this, the output dimension of the Spatial Attention (SA) module remains unchanged, and its output is upsampled to the nearest neighbor in the time dimension. The time length T is extended to αT, which aligns it with the fast path in the temporal dimension. Finally, the output is incorporated into the features of the fast path, illustrated as

$$Out_f^{[B,\beta C+\beta C,\alpha T,H,W]} = Con\left(SA\left(Down(I_s^{[B,C,T,H,W]})\right), I_f^{[B,\beta C,\alpha T,H,W]}\right) \quad (1)$$

where Down($\cdot$) is the down-sampling operation, Con($\cdot$) is the fusion operation in the given dimension, and SA($\cdot$) is the spatial attention module. In the proposed model, the extended CA^{3D} (3D-Coordinate Attention) is used as the spatial attention module.

The extended CA^{3D} definition is shown in Fig 2. Based on the CA module proposed by Hou et al. [25], its temporal dimension is extended as below. For a given input $X = [x_1, x_2 \ldots, x_c]$ three pooling kernels of size $(1, W, T)$, $(H, 1, T)$ and $(H, W, 1)$ are used along the horizontal, vertical and temporal coordinates, respectively. For each channel encoded to obtain z, it is encoded as

$$z_h^c(h) = \frac{1}{WT} \sum_{\substack{0 \le j < W \\ 0 \le g < T}} x_c\,(h, j, g) \quad (2)$$

$$z_w^c(w) = \frac{1}{HT} \sum_{\substack{0 \le i < H \\ 0 \le g < T}} x_c\,(i, w, g) \quad (3)$$

$$z_t^c(t) = \frac{1}{HW} \sum_{\substack{0 \le i < H \\ 0 \le j < W}} x_c\,(i, j, t) \quad (4)$$

The three transformations described above enable feature aggregation in both spatial and temporal directions, resulting in a pair of spatiotemporal perceptual feature maps. This approach allows our attention module to capture area block features in both the two-dimensional spatial and temporal directions, respectively, which enhances the network's ability to accurately locate the features of interest. After obtaining the corresponding feature maps separately, followed by a shared $1 \times 1 \times 1$ convolutional transformation, feature fusion is performed by

$$f = \delta(F_1\left([z_h^c, z_w^c, z_t^c]\right)) \quad (5)$$

where the function $F_1(\cdot)$ is a shared $1 \times 1 \times 1$ convolutional transform function, $[\cdot, \cdot]$ is a concat operation along the spatial dimension, and δ is a nonlinear activation function.

After the above operation, the feature map $f \in \mathbb{R}^{\frac{C}{r}\times(H+W+T)}$ is obtained. The feature map f is decomposed in the spatial dimension to obtain three independent tensors $f^t = \mathbb{R}^{\frac{C}{r}\times T}$, $f^h = \mathbb{R}^{\frac{C}{r}\times H}$ and $f^w = \mathbb{R}^{\frac{C}{r}\times W}$. Then, using the three convolutional transformations, it will be converted into a tensor with the same channel size as the input X. Finally, the corresponding attention weights $d_c^t\,(g)$, $d_c^h(i)$ and $d_c^w(j)$

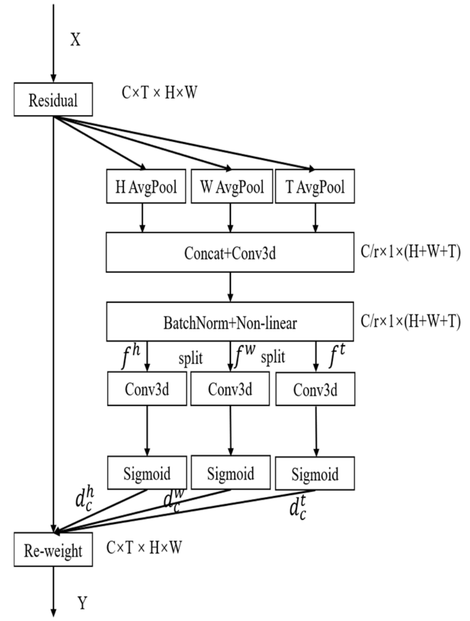

FIGURE 2. 3D coordinate attention module.

obtained after the sigmoid function will be subjected to a matrix product operation with the original input to obtain $y_c(g, i, j)$:

$$y_c(i, j, g) = y_c(i, j, g) \times d_c^h(i) \times d_c^w(j) \times d_c^t(g) \quad (6)$$

B. CHANNEL ATTENTION

There are two functions of the channel attention (CA), including:

-Adaptively adjust the weight of each fast path to the slow path to improve the model's attention to different channels;

- Extract important features in the input data and focus more attention on these features to better capture important information in the input data and improve the performance of the model.

The feature fusion of channel attention is given by

$$\begin{aligned} &Out_s^{[B,\beta C+C,T,H,W]} \\ &= Con\left(CA\left(Down(I_f^{[B,\beta C,\alpha T,H,W]})\right), I_s^{[B,C,T,H,W]}\right) \quad (7) \end{aligned}$$

where $I_f^{[B,\beta C,\alpha T,H,W]}$ is the fast path initial input, Down(·) is the maximum pooling downsampling in the T -th dimension, $I_s^{[B,C,T,H,W]}$ is the slow path initial input, Con(·) is the fusion operation in the specified dimension, and CA(·) is the channel attention module. For the specific fusion, maximum pooling is chosen to downsample the fast path in the T dimension after the output of the CA module matches the input dimension. Finally the output of the CA module is fused with the slow path in the channel dimension for feature information.

For the CA module in the present model, we chose CAM module proposed by Woo et al. [26]. The corresponding module is shown in Fig. 3 The overall architecture of the CAM module is similar to SENet [27], except that it uses both the average pool feature F_{avg}^c and the maximum pool feature F_{max}^c. These two features are used to greatly improve the representational ability of the network. A Multi-Layer Perceptron MLP is used to compute the importance weights $M_c(F)$ for each channel, depicted as

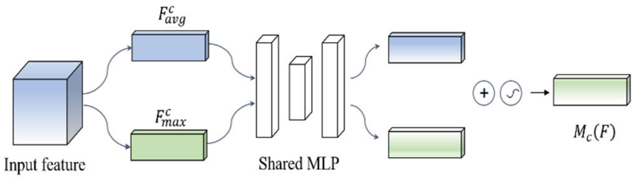

FIGURE 3. CAM module.

FIGURE 4. Selected UCF101 records.

$$\begin{aligned} M_c(F) &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0\left(F_{max}^c\right)\right)\right) \quad (8) \end{aligned}$$

where σ is a sigmoid function, W_0 and W_1 share MLP weights, $W_0 \in \mathbb{R}^{\frac{C}{r} \times C}$, and $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$.

IV. BENCHMARK DATASET EXPERIMENTS

A. EXPERIMENTS SETUP

The effectiveness of the proposed model is first validated using the publicly available behavior dataset UCF101 [28]. This dataset, derived from the YouTube video website, is a typical video dataset for behavior recognition. It consists of a total of 101 action categories, 13,320 videos, and 27 hours of footage. The dataset includes five main action categories: (1) human-object interaction, (2) simple physical action, (3) human-human interaction, (4) playing a musical instrument, and (5) sports. Fig. 4 displays some of the behaviors included in the dataset.

The experimental hardware utilized is the Inspur Yingxin server NF5280M6, equipped with the NVIDIA A40 GPU

TABLE 1. Network hyperparameter settings.

Parameter	Value	Description
EPOCH	120	Maximum number of training iterations
BATCH_SIZE	16	Training sample size per batch
NUM_FRAMES	32	Number of frames to sample
SAMPLING_RATE	2	Frame sampling rate (Interval Between Two Sampled Frames)
OPTIMIZER	SGD	Stochastic gradient descent MOMENTUM=0.9
LOSS_FUNC	cross_entropy	Cross entropy loss
LR	0.1	Initial value of the learning rate
LR_POLICY	steps_with_relative_lrs	Dynamic adjustment strategies for learning rates STEPS: [0, 50,70, 90] LRS: [1, 0.1, 0.01, 0.001]
α	8	The frame rate ratio between the Fast and Slow Pathways
β	1/8	The channel ratio between the Fast and Slow pathways
RNG_SEED	42	Ensures reproducible results and also helps to improve the stability and reliability of the model

TABLE 2. Evaluation indicators.

Indicators	Function
Precision	The number of positive class samples predicted by the classifier as a proportion of the number of all positive class samples predicted by the classifier, which measures the accuracy of the classifier
Recall	A measure of the completeness of the classifier: the number of positive class samples predicted by the classifier as a proportion of the number of all positive class samples
F1-Score	A comprehensive measure of classifier performance, factoring in the effects of accuracy and recall, to help evaluate classifiers more comprehensively
Params	The number of parameters refers to the total number of parameters to be trained in the network model
Top-1	Indicates that the category with the greatest probability in the category probability is exactly the probability of the object category
FLOPs	Refers to the number of floating point operations, a measure of the computational power of a network model 1GFLOPs=10^9FLOPs

graphics card. The software environment employed is the Ubuntu 18.04 LTS operating system. All evaluated deep learning models are based on PyTorch (version 1.13.0). An important objective of this work is to conduct ablation experiments to confirm the validity of the models. Since fixed model parameters are necessary for ablation experiments, the hyperparameters were uniformly set for all models, as presented in Table 1.

To make the model training more appropriate, the experiments used the Warm_up pre-warming training strategy. The initial learning rate for pre-warming was 0.01 and the total number of pre-warmings was 34. Dropout regularisation was used to mitigate overfitting problems and set Dropout_rate to 0.5. The training, validation, and test datasets were split 6:2:2. The input image size was [704,576], and the input image was randomly cropped to size [224,224] in training and [256,256] in testing.

For evaluating the proposed method, the experiments evaluation metrics as shown in Table 2 were used in the experiments.

B. ABLATION EXPERIMENT

The proposed SlowFast network with spatio-temporal attention fusion comprises three essential components: spatial attention, channel attention, and attention fusion. This section details the mechanisms for acquiring spatial and channel attention, along with the attentional fusion mechanism. It also employs ablation experiments to select the optimal model.

Regarding channel attention, we conducted a comparative evaluation between the CAM module and the ECA module [29], which is an enhancement of SENet [27]. Although SENet's dimensionality reduction reduces model complexity, it breaks the direct correspondence between channels and their weights. Therefore, Wang et al. [29] decided not to employ dimensionality reduction to calculate channel attention, instead trading a small number of parameters for significant performance improvement.

Concerning spatial attention, we comparatively evaluated three attentional mechanisms: Coordinate Attention, Spatio-Temporal Attention (ST-Attention) [30], and Extended 3D Coordinate Attention (CA^{3D}). We selected ST-Attention, which extends self-attention along the temporal axis to a 3D temporal convolution module, as proposed by Wei et al. [30]. The module is applicable to the video domain and can be directly incorporated into many other spatiotemporal networks, enabling the model to accurately capture dynamic changes and spatiotemporal structure in the video, thus enhancing the model's performance in these tasks. For the base model, the following three options were considered. (1) Scheme 1 investigates the impact of channel attention on the model. We employed the basic SlowFast model as a foundation and substituted the convolutional fusion connection method with the Channel Attention (CA) module. (2) Scheme 2: To investigate the effect of using convolution for fusion connection in different single pathways, the SlowFast model was used as the base and the fusion direction was changed.

TABLE 3. Results of ablation experiments.

Base Model	Fast fusion style			Slow fusion style				Params	Top-1	GFLOPs
	Conv	ECA	CAM	Conv	STA	Coo	CA^{3D}			
SlowFast	√							33.79M	95.93%	36.336
SlowFast		√						32.98M	96.41%	35.107
SlowFast			√					32.98M	95.89%	35.107
SlowFast				√				32.63M	95.34%	35.668
SlowFast					√			32.70M	94.59%	59.714
SlowFast						√		32.64M	96.22%	35.668
SlowFast							√	32.64M	96.63%	35.669
SlowFast	√			√				34.04M	95.93%	37.613
SlowFast	√				√			34.11M	93.37%	61.658
SlowFast	√					√		34.05M	95.78%	37.614
SlowFast	√						√	34.05M	95.56%	37.613
SlowFast		√			√			33.29M	95.41%	60.409
SlowFast		√				√		33.23M	97.04%	36.385
SlowFast		√					√	33.24M	96.52%	36.384
SlowFast			√		√			33.30M	96.15%	60.429
SlowFast			√			√		33.24M	97.33%	36.385
SlowFast			√				√	33.24M	**97.45%**	36.384

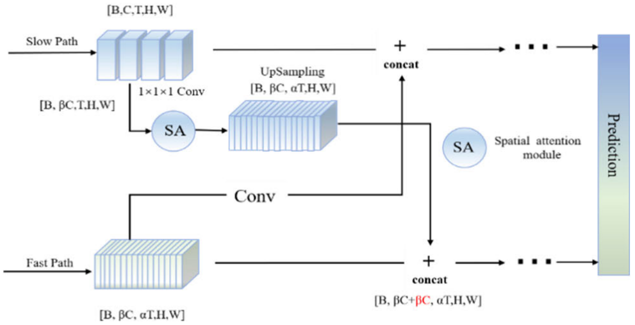

FIGURE 5. Converged connection network.

Only the SA module was used for fusion. (3) Scheme 3 investigates the impact of adding Slow-Fast paths and how different spatial attention mechanisms affect the network. We employed the Slow-Fast model as the base and introduced a Slow-Fast path connecting to the SA module. This retained the original Fast path and Slow path, which were fused through convolutional connection. The model's architecture is illustrated in Fig. 5. (4) Scheme 4 explores the the impact of combining different types of channel attention and spatial attention. We modified the original convolutional fusion method of the Fast path in Fig. 5 by incorporating the CA module for fusion connection. Fig. 1 depicts the model's structure, with both the SA module and CA module altered.

A total of 16 combinations were considered, as presented in Table 3, which also displays the experimental results of the different combinations. From the results of Scheme 1 and Scheme 3, it is evident that merely changing the convolutional fusion method of the original SlowFast model does not significantly enhance, or even slightly decreases, the recognition performance. Meanwhile, the experimental results of Scheme 2 demonstrate that changing the fusion direction alone has little impact on the model. The same is true when adding the SA module without changing the convolutional fusion of the Fast path. However, fusing the three combinations of channel attention and spatial attention leads to a significant boost in model performance, with a small computational cost for a considerable increase in accuracy. Specifically, SlowFast+ CA^{3D} +CAM achieved a 1.52% improvement in Top-1 accuracy.

C. COMPARATIVE EXPERIMENT

As presented in Table 3, the combination of SlowFast+ CA^{3D} +CAM yielded the best results on the UCF101 dataset, thus selected as the final proposed scheme for comparison with four other models, namely C3D, Timesformer, TSN [31] and basic SlowFast. Table 4 shows the experimental results, where the proposed scheme only slightly falls short of the Top-1 value of Timesformer. However, when it comes to GFLOPs, which reflects computational power, Timesformer's computational complexity is 30 times higher than the proposed solution. Furthermore, STAF-SlowFast has fewer model parameters than C3D, Timesformer, and SlowFast, and only slightly more than TSN, which has the lowest number of parameters but only marginally better performance than C3D. STAF-SlowFast outperforms the other four models across all three metrics (number of model parameters, Top-1 accuracy, and GFLOPs).

D. ATTENTIONAL FUSION MECHANISMS ANALYSIS

To investigate the contribution of attentional fusion mechanisms in the network, we used a Grad-CAM visualisation

TABLE 4. **Comparison of performance of existing models.**

Model	Params	Top-1	GFLOPs
C3D [11]	78.02M	93.71%	38.615
Timesformer [15]	115.64M	98.12%	1036.444
TSN [31]	22.43M	94.00%	32.959
SlowFast [12]	33.79M	95.93%	36.336
STAF-SlowFast	33.24M	**97.45%**	36.384

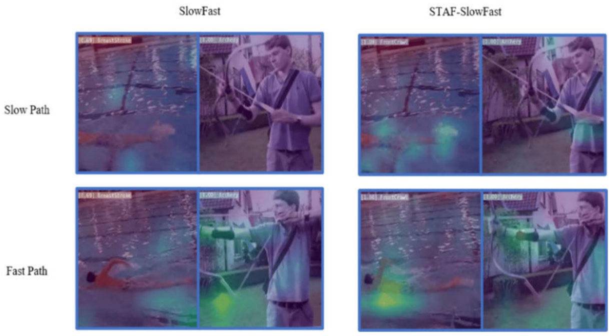

FIGURE 6. **Map of Grad-CAM visualization features.**

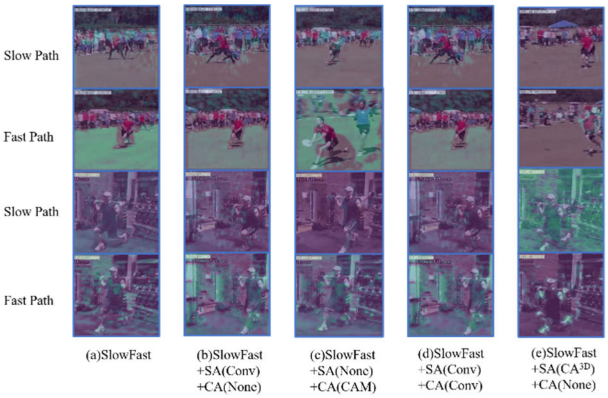
Slow Path
Fast Path
Slow Path
Fast Path
(a)SlowFast
(b)SlowFast +SA(Conv) +CA(None)
(c)SlowFast +SA(None) +CA(CAM)
(d)SlowFast +SA(Conv) +CA(Conv)
(e)SlowFast +SA(CA3D) +CA(None)

FIGURE 7. **Comparison of fusion convolution module and attention module heat maps.**

TABLE 5. **STAF-SlowFast comparative results with other models.**

Model	Params	Top-1	GFLOPs
C3D [11]	78.02M	51.60%	38.615
Timesformer [15]	115.64M	78.54%	1036.444
TSN [31]	22.43M	69.40%	32.959
SlowFast [12]	33.57M	67.30%	36.336
STFA-SlowFast	33.02M	69.85%	36.384

model [32] for evaluation. We viewed the SlowFast and STAF-SlowFast heat maps for the two channel convolutional layers prior to output. We selected two videos from UCF101 for the heat map visualisation analysis, and the final results are shown in Fig. 6. The figure clearly shows that the original SlowFast network does not focus well on the key feature regions, whether in the slow or fast path. The proposed STAF-SlowFast network mainly focuses on the human body structure in the Slow path, and the Fast path mainly focuses on the video motion pixel part, which helps the network to perform spatiotemporal feature extraction.

For the role of different selections of SA and CA modules corresponding to the model, the Grad-Cam visualization model was used to observe the heat map after the first fusion behavior of the model.

Heat maps (a) and (b) in Fig. 7 show that the two separate paths using convolution for information fusion have similar effects, as evidenced by the uniform heat distribution on the surface of the video in both the slow and fast paths. This proves that the author's choice of fusion direction is correct and that the flow direction of the fusion does not make a practical difference to the model. Comparing (a) and (c), it can be observed that the model using attention modules exhibits a more concentrated heat map in the spatial dimension, which is beneficial for extracting spatial backbone information. When comparing (a) and (d), it is evident that the distribution of the heat map is relatively uniform. This suggests that the convolutional fusion of the two paths does not significantly enhance the extraction of spatiotemporal information compared to the original model. The heat map (d) with the CA^{3D} module exhibits a more concentrated heat map in the temporal dimension, which is beneficial for extracting temporal backbone information. In summary, the SA and CA modules using attention mechanisms are more effective for extracting spatiotemporal information than traditional convolution modules, and fusing information can effectively help the network to train feature extraction.

E. HMDB51 DATASET'S-BASED EVALUATION

From the experimental results, it can be observed that for datasets with large amounts and diverse categories of data, our proposed model can outperform the original SlowFast network in terms of performance. However, it is unclear how our model performs on datasets with small amounts and few categories of data. Therefore, we investigated the performance of our model using the HMDB51 dataset [33].

The HMDB51 dataset consists mostly of clips from movies, as well as a small portion from public databases such as the Prelinger Archive, YouTube, and Google Videos. The dataset contains 6849 clips, divided into 51 action categories, with at least 101 clips per category. Action categories can be classified into five types: (1) facial actions without object manipulation, (2) facial actions with object manipulation, (3) general body movements, (4) body movements with object interactions, and (5) human interactions with each other.

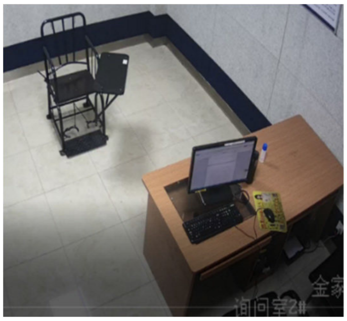

FIGURE 8. Environment of the interrogation room.

According to Table 5, it can be observed that our proposed model outperforms all models except Timesformer on the small dataset. Due to the limited amount of data and the simplicity of the model, the Top-1 metric is lower than that of Timesformer. However, both the number of parameters (Params) and the number of GFLOPs are significantly lower than Timesformer. In summary, the proposed STAF-SlowFast improvement is effective, as it demonstrates performance improvements on both small and large datasets.

V. APPLICATION FOR IDENTIFYING VIOLATIONS

This work is driven by the actual requirements of a Chinese public security agency. In this section, we apply the proposed system to detect interrogation violations. In consideration of the sensitivity of the data and the protection of personal privacy, the dataset used for training is not directly obtained from the interrogation scene surveillance video. Instead, it is a simulation and emulation of the interrogation violation scenario under the supervision of a professional judge, who then records the relevant violations.

In order to detect abnormal behavior while also distinguishing between normal and abnormal behaviors, we constructed two datasets, namely Dataset I and Dataset II. Dataset I consists of various categories of abnormal behaviors and solely simulates the occurrence of misconduct within interrogation rooms. It is used to evaluate the model's ability to accurately identify abnormal behaviors. The purpose of constructing Dataset II is to provide a more realistic dataset that better simulates the interrogation room environment and the occurrence of unforeseen events. Dataset II is an extension of Dataset I and includes additional categories of abnormal behaviors as well as normal behaviors. The model is expected to recognize both types of behaviors and accurately identify different types of abnormal behaviors within Dataset II.

A. DATA ACQUISITION

The simulated interrogation room environment is depicted in Fig. 8, with the camera positioned above the interrogator's seat. The camera model used is a Hikon camera DS-2DC4423IW-D, featuring 4 megapixels, a maximum aperture of F1.6, a focal length ranging from 4.8mm to 110mm, and a video recording size of 704 × 576 pixels.

TABLE 6. Four categories of behaviors and the determining rules.

Class	Conduct rules
Assault	Assault on the person being interrogated, specific acts of violence are punches, slaps, kicks, tearing etc.
Long Squat	Requiring the person to crouch for more than three seconds is considered a long squat.
Push-up	Interrogation room personnel performing push-ups and multiple repetitions of the action
Abnormal Running	Interrogators running around the interrogation room for more than three seconds is considered an abnormal run.

TABLE 7. Expansion behavior and rules.

Class	Conduct rules
Abnormal Jumping	Personnel in the interrogation room engaging in jumping behavior for more than three seconds is considered abnormal jumping
Assault With Stick	Use of weapons such as clubs by persons in the interrogation room to physically harm other persons
Normal Behavior	The personnel in the interrogation room carried out their activities normally and no irregularities were identified

The Dataset I was constructed by categorizing small video clips and storing them in folders based on specific rules. All clips were saved in .avi format. Table 6 shows the categories and the corresponding rules used for categorization.

Three new categories, namely abnormal jumping, fighting, and normal behavior, were added to Dataset II. Normal behavior was used as a control group for abnormal behavior. The specific descriptions of the newly added behaviors are shown in Table 7.

To increase the complexity of the dataset and better simulate real-life scenarios, stools, tables, and people were added to block the view of the person committing the act during the experiment. During filming, the number of people gradually increased from one to four, adjustments were made to the brightness of the location, changes were made to clothing and the use of hats, and the camera angle was varied to subject the model training to a range of factors. For the dataset, to enhance its challenging nature, certain portions of the collected data are subjected to operations such as rotation, flipping, adding noise, and applying color filters. This aims to improve the generalization capability and robustness of the learning model in later stages.

Following the shooting scheme designed above, the details of dataset I are obtained as shown in Table 8.

TABLE 8. Interrogation behavior dataset I.

Class	Num	Average Duration
Assault	190	2.5s
Long Squat	272	3s
Push-Up	183	3s
Abnormal Running	267	2.5s

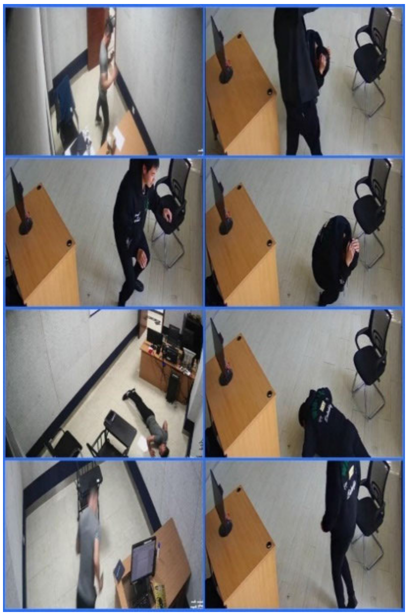

FIGURE 9. Selected trial infringement datasets.

TABLE 9. Interrogation behavior dataset II.

Class	Num	Average Duration
Assault	397	3.78s
Long Squat	404	4.23s
Push-Up	287	3.21s
Abnormal Jumping	287	3.47s
Abnormal Running	356	3.59s
Assault With Stick	291	3.54s
Normal	353	4.28s

The Dataset I comprises videos of the real environment and the simulated construction environment and includes four action categories: Assault, Long Squat, Push-Up, and Abnormal Running. The total number of videos is 912 with a total duration of 3 hours and 16 minutes. Fig. 9 displays some of the recorded action clips.

Dataset II is more challenging as compared to Dataset I with larger amount of data and more categories. The detailed description of the collected datasets is shown in Table 9.

Dataset II comprises videos from both real-world scenarios and simulated environments. It consists of a total of seven action categories, namely, assault, long squat, push-ups, abnormal running, abnormal jumping, assault with stick, and normal behavior group. The dataset comprises 2375 video clips, with a combined duration of 5 hours and 8 minutes. Among them, Fig. 10 shows some newly added action segments captured during filming.

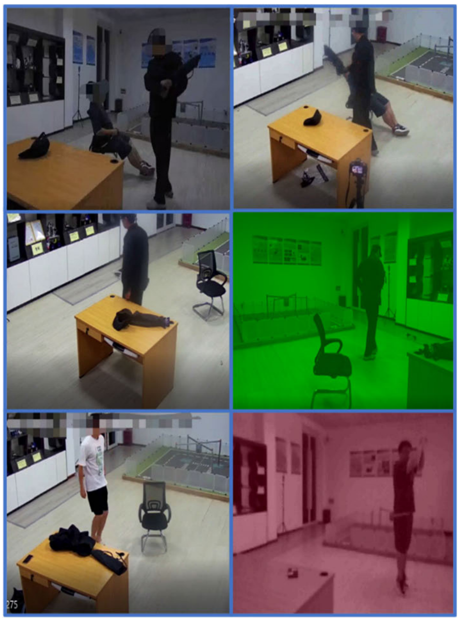

FIGURE 10. Some of the new action category fragments.

TABLE 10. Results of the interrogation dataset I for different models.

Module	Params	Top-1	GFLOPs
C3D [11]	78.02M	96.57%	38.615
Timesformer [15]	115.64M	97.28%	1036.444
TSN [31]	22.43M	92.93%	32.959
SlowFast [12]	33.57M	97.28%	36.336
STAF-SlowFast	33.02M	**98.91%**	36.384

B. PERFORMANCE EVALUATION BASED ON DATASET I

All metrics listed in Table 1 were used for evaluation, and the model training parameters remained the same as in Section IV-A (Table 2). Following the ablation experiments on the public dataset, our proposed STAF-SlowFast architecture exhibited superior performance. Table 10 compares the results of various models with our proposed model on the interrogation violations dataset. The correct rates for each network model were 96.57% (C3D), 97.28% (Timesformer), 97.28% (SlowFast), 92.93% (TSN), and 98.91% (STAF-SlowFast), respectively.

The results above demonstrate that the STAF-SlowFast network outperforms other existing behavior recognition models on the interrogation violations dataset, achieving the best results in the Top-1 metric. Table 11 presents the results

TABLE 11. STAF model performance compared to other existing models on Dataset I.

Model	Assault			Long Squat			Push-Up			Abnormal Running		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
C3D	98.00%	96.08%	90.01%	95.24%	90.91%	93.02%	93.33%	92.11%	68.65%	98.21%	100.00%	99.10%
Timesformer	100.00%	100.00%	100.00%	95.00%	93.14%	94.06%	97.00%	100.00%	98.48%	96.00%	95.05%	95.52%
TSN	100.00%	96.43%	98.18%	91.89%	85.00%	88.31%	89.47%	94.44%	91.89%	89.09%	94.23%	91.59%
SlowFast	100.00%	100.00%	100.00%	92.00%	100.00%	95.83%	97.00%	93.27%	95.10%	98.00%	94.23%	96.08%
STAF-SlowFast	98.00%	100.00%	98.99%	100.00%	100.00%	100.00%	100.00%	96.15%	98.04%	98.00%	100.00%	98.99%

TABLE 12. Comparison of STAF-SlowFast with existing models on Dataset II.

Model	Evaluation Indicators			Assault			Long Squat			Push-Up		
	Params	Top-1	GFLOPs	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
C3D	82.12M	92.78%	38.615	98.00%	96.08%	90.01%	95.24%	90.91%	93.02%	93.33%	92.11%	68.65%
Timesformer	118.53M	97.49%	1036.444	100.00%	100.00%	100.00%	95.00%	93.14%	94.06%	97.00%	100.00%	98.48%
TSN	22.43M	94.75%	32.959	100.00%	96.43%	98.18%	91.89%	85.00%	88.31%	89.47%	94.44%	91.89%
SlowFast	33.57M	98.54%	36.336	100.00%	100.00%	100.00%	92.00%	100.00%	95.83%	97.00%	93.27%	95.10%
STAF-SlowFast	33.02M	99.16%	36.384	98.00%	100.00%	98.99%	100.00%	100.00%	100.00%	100.00%	96.15%	98.04%

Model	Abnormal Running			Abnormal Jumping			Assault With Stick			Normal		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
C3D	98.21%	100.00%	99.10%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	93.42%	100.00%	96.60%
Timesformer	96.00%	95.05%	95.52%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.00%	99.50%
TSN	89.09%	94.23%	91.59%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	91.29%	95.45%
SlowFast	98.00%	94.23%	96.08%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.00%	99.50%
STAF-SlowFast	98.00%	100.00%	98.99%	100.00%	100.00%	100.00%	100.00%	99.00%	99.50%	100.00%	99.00%	99.50%

of test sample identification for the different models and the STAF-SlowFast network separately.

According to Table 11, it can be observed that the C3D, Timesformer, TSN, and SlowFast models show lower detection metrics for long squats, push-ups, and abnormal running behaviors, indicating their inability to effectively detect these behaviors. However, a significant improvement in accuracy is observed for the STAF-SlowFast in all behavior categories, indicating its ability to accurately identify them. The recall rate for all three categories reached 100%, indicating that the STAF-SlowFast network can capture the characteristics of each category very well. The F1-score evaluation metrics also showed improvements compared to other models, demonstrating the effectiveness of our network model in identifying and classifying irregular violations.

C. PERFORMANCE EVALUATION BASED ON DATASET II

The model training parameters were uniformly applied for performance evaluation using the metrics from Dataset I. In Table 12, we compared the results of different models with our proposed model on the newly constructed interrogation behavior dataset. The accuracy of each network model was as follows: 92.78% (C3D), 97.49% (Timesformer), 98.54% (SlowFast), 94.75% (TSN), and 99.16% (STAF-SlowFast).

The column 'Evaluation Indicators' in the table 12 demonstrates that the STAF-SlowFast network outperforms other existing behavior recognition models on the new interrogation misconduct dataset, achieving the best performance in the top-1 metric. Due to the considerably larger size of the new dataset compared to the initial one, the final results are higher than those obtained on the initial dataset. To investigate the advantages of the STAF-SlowFast over other models, Table 12 presents the recognition results of different models compared to STAF-SlowFast on the test samples.

In the remaining part of Table 12, our proposed STAF-SlowFast demonstrates excellent performance across all categories. As interrogation misconduct may pose a threat to the safety of personnel inside the interrogation room, our model needs to accurately identify each misconduct behavior, respond promptly, and distinguish between normal and abnormal behaviors. By observing the table, it is evident that the STAF-SlowFast, with the inclusion of the pathway attention mechanism, outperforms the original SlowFast in all three metrics. This indirectly demonstrates the feasibility and enhanced accuracy of our approach in identifying violation behaviors.

VI. CONCLUSION

This paper proposes a spatio-temporal attention fusion SlowFast network for intelligent detection of interrogation violations. The model utilizes the slow-fast and fast-slow pathways to exchange information between different layers and introduces an attention mechanism to focus on the regions and temporal sequences required for action recognition. Through ablation experiments, the attention acquisition and fusion mechanisms in the spatial and temporal channels are optimized. The combination of Slow-Fast + CA^{3D} + CAM achieves the best performance. On the UCF101 dataset, the proposed model improves the Top-1 detection accuracy by 1.52% compared to the SlowFast. On the HMDB51 dataset, the accuracy reaches 69.85%, surpassing the SlowFast but still leaving room for further improvement. Furthermore, we further apply this model to the identification of interrogation misconduct in Dataset I and Dataset II, achieving accuracies of 98.91% and 99.16% respectively. This demonstrates the effectiveness of the model in recognizing indoor misconduct behaviors during interrogations. Compared to some state-of-the-art behavior recognition models, the proposed model proves to be highly competitive. However, it still faces challenges in training on insufficient data, such as the HMDB51 dataset. In future work, we will focus on addressing this issue.

REFERENCES

[1] X. Wang, "Research on detection method of human abnormal behavior in micro-geographic environment—Take interrogation room for example," M.S. thesis, Dept. Geographical Sci., Nanjing Normal Univ., Nanjing, China, 2013.

[2] J. Li, K. Shang, and H. Liu, "Research on the law enforcement supervision system of intelligent detection of non-standard behaviors of smart prosecutors," in *Proc. Beijing Justice Network Media Conf.*, 2019, pp. 226–245.

[3] D. Tan, W. Wang, and Y. Wang, "Recognition and detection of violence in public security surveillance video," in *J. People's Public Secur. Univ. China*, vol. 27, no. 2, pp. 94–100, 2021.

[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE 17th Int. Conf. Pattern Recognit.*, vol. 3, Aug. 2004, pp. 32–36.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[6] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 357–360.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional twostream network fusion for video action recognition," in *Proc. CVPR*, Jun. 2016, pp. 1933–1941.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.

[13] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[14] H. Ge, Z. Yan, W. Yu, and L. Sun, "An attention mechanism based convolutional LSTM network for video action recognition," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 20533–20556, Jul. 2019.

[15] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021, pp. 1–12.

[16] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Adv. Neural Inf. Process. Syst. Workshops*, Dec. 2015, pp. 1–4.

[17] M. Patrick, D. Campbell, Y. M. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12493–12506.

[18] W. Xiang, C. Li, B. Wang, X. Wei, X.-S. Hua, and L. Zhang, "Spatiotemporal self-attention modeling with temporal patch shift for action recognition," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, Oct. 2022, pp. 627–644.

[19] A. Datta, M. Shah, and N. D. V. Lobo, "Person-on-person violence detection in video data," in *Proc. Int. Conf. Pattern Recognit.*, vol. 1, Aug. 2002, pp. 433–438.

[20] V. Lam, S. Phan, D.-D. Le, D. A. Duong, and S. Satoh, "Evaluation of multiple features for violent scenes detection," *Multimedia Tools Appl.*, vol. 76, pp. 7041–7065, Mar. 2017.

[21] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent Flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–6.

[22] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," in *Proc. Int. Symp. Vis. Comput.*, 2014, pp. 551–558.

[23] Z. Dong, J. Qin, and Y. Wang, "Multi-stream deep networks for person to person violence detection in videos," in *Proc. 7th Chin. Conf. Pattern Recognit. (CCPR)*. Chengdu, China: Springer, Nov. 2016, pp. 517–531.

[24] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," *J. Phys., Conf. Ser.*, vol. 844, no. 1, Apr. 2017, Art. no. 012044.

[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[26] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 3–19.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[28] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[30] D. Wei, Y. Tian, L. Wei, H. Zhong, S. Chen, S. Pu, and H. Lu, "Efficient dual attention SlowFast networks for video action recognition," *Comput. Vis. Image Understand.*, vol. 222, Sep. 2022, Art. no. 103484.

[31] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2016, pp. 20–36.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[33] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. ICCV*, Nov. 2011, pp. 2556–2563.

HAILUN WANG received the B.Sc. degree in automotive engineering from the Changchun University of Technology, China, in 2002, the M.Sc. degree in electronics and information engineering from the Zhejiang University of Technology, China, in 2006, and the Ph.D. degree in power electronics and power transmission from Shanghai Maritime University, in 2021. She is currently the Vice President and a Professor with the School of Electrical and Information Engineering, Quzhou University. Her research interests include information fusion, nonlinear filter, target tracking, fault diagnosis of hydropower, and power transmission and distribution equipment.

BIN DONG received the B.S. degree from the Anhui University of Science and Technology, China, in 2021. He is currently pursuing the master's degree with Hangzhou Dianzi University. His research interest includes video behavior recognition.

QIRUI ZHU received the B.S. degree from the Anhui University of Finance and Economics, China, in 2021. He is currently pursuing the master's degree with Hangzhou Dianzi University. His research interest includes leather defect detection.

ZHIQIANG CHEN received the B.S. degree from the Wuhan University of Water-Conservancy and Electric Power, Wuhan, China, in 2001, the M.S. degree from Chongqing University, Chongqing, China, in 2004, and the Ph.D. degree from the University of Fukui, Japan, in 2011. He is currently a Professor with the School of Electrical and Information Engineering, Quzhou University. His research interests include data mining, machine vision-based surface defect detection and prognostics, and health management.

YI CHEN (Senior Member, IEEE) received the B.Sc. degree in automotive engineering from the Chongqing University of Technology, in 2001, the M.Sc. degree in automotive engineering from Chongqing University, in 2004, and the Ph.D. degree in mechanical engineering from the University of Glasgow, in 2010. He has been taking a leading role in the previous and current department to maintain cross-disciplinary research links, and develop external research collaborations both nationally and internationally. He has been leading a few research grants in the areas of artificial intelligence, high-performance computing, robotics and autonomous systems, and also studies in multi-disciplinary contexts. He is currently a Chartered Engineer. He has published more than 100 academic papers in both high impact international academic journal and international conferences and has been selected as a Publons' top 1% of reviewers in computer science and engineering. He has been actively involved in both academic research and KTP projects, as a PI and a CoI, funded by EPSRC, U.K.; Horizon2020, EU; NSFC, China; National Key Research and Development Program of China; and Industrial funding bodies. He is a member of AAAI, AIAA, and ASME, and a fellow of HEA, RSA, IET, and IMechE.