STFE-Net:Enhanced Deep Learning for Real-Time Abnormal Behavior Detection in Video Surveillance

Cong Chen [1], Xianjun Fu [2], Yi Li [3*]
[1]Computer Public Training Center, Zhejiang College of Security Technology; Wenzhou, Zhejiang, 325035, China
[2]Institute for Artificial Intelligence, Zhejiang College of Security Technology; Wenzhou, Zhejiang, 325035, China
[3]College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, Zhejiang, 325035, China
E-mail：liy147852@outlook.com
[*]Corresponding author

Keywords: Deep learning, real-time video surveillance, abnormal behavior detection, spatiotemporal features, STFE-Net

Received: July 11, 2025

A spatiotemporal feature enhancement network (STFE-Net) is proposed for real-time abnormal behavior detection in surveillance video. The model integrates optimized 3D convolution kernels, a multi-scale convolution strategy, and a feature fusion module with skip connections and attention mechanisms, followed by an improved fully connected classification structure. STFE-Net achieves an average accuracy of 0.85 on the UCF-Crime dataset and 0.82 on multiple datasets, outperforming traditional 3D-CNNs (average 0.65) and RNN-based models (average 0.68). False alarm rates are reduced to 0.10 and false negative rates to 0.07, demonstrating improved precision and robustness. Compared with baseline methods, STFE-Net shows a 26.2% increase in average accuracy and up to 50% reduction in false positives, significantly improving real-time surveillance reliability.

Povzetek: Študija predstavi STFE-Net, model za zaznavanje nenormalnega vedenja v nadzornih videih, ki z optimiziranimi 3D jedri, konvolucijo ter pozornostjo izboljša natančnost in občutno zmanjša lažne alarme glede na klasične 3D-CNN/RNN pristope.

1 Introduction

In today's society, security issues have become the core of people's concern. Various places, such as shopping malls, schools, streets, etc., are in urgent need of effective monitoring methods to ensure public safety. According to incomplete statistics, there are as many as 100 million security accidents caused by various abnormal behaviors every year in the world, causing economic losses of more than 500 billion US dollars. Among them, a considerable number of accidents can be completely avoided or reduced if the abnormal behaviors can be detected and intervened in time with the help of effective video monitoring methods in the early stage. Take a large city as an example. In the past year, it has carried out spot checks on video surveillance in public places. The results show that about 30% of the monitoring time periods have monitoring loopholes, that is, potential abnormal behaviors such as theft and Figurehting cannot be detected in time. This fully demonstrates that the existing video surveillance system has serious deficiencies in abnormal behavior detection.

Recent literature has demonstrated rapid progress in deep learning-based anomaly detection for video surveillance. Umale-Nagmote et al. [1] proposed a hybrid model combining spatiotemporal autoencoders and convolutional LSTMs, enabling effective representation of complex behavior dynamics. Aberkane and Elarbi-Boudihir [2] introduced a reinforcement learning-based framework that adaptively detects anomalies through environment-agent interactions, supporting real-time responsiveness. Alarcon et al. [3], while focusing on distribution network planning, highlighted the value of capacity-efficient modeling strategies, indirectly informing resource-aware deep learning deployment in real-time systems. Zhang et al. [4] developed an online anomaly detection method emphasizing fast response and stream adaptability, crucial for continuous surveillance operations. Chen et al. [5] advanced 3D dense connection architectures to improve the representation of spatial-temporal behavior correlations, reducing misclassification in complex scenes. Li et al. [6] contributed robust planning mechanisms based on correlation modeling, offering parallels in designing neural structures sensitive to contextual dependencies. Chen and Zhang [7] applied involution feature extraction to human posture analysis, which aligns with the need for refined motion detail in anomaly recognition.

Some studies use manual feature extraction combined with machine learning classifiers, such as optical flow features, HOG features, etc., to classify abnormal behaviors with the help of classifiers such as support vector machines and random forests. However, these manual features often have limitations and cannot fully capture the complex spatiotemporal information in the video. In complex scenes, the accuracy of abnormal behavior detection is relatively limited. In addition, methods based on deep learning have also emerged, such as the use of convolutional neural networks (CNN). Some studies have achieved relatively high detection accuracy in certain specific scenarios by constructing a multi-layer CNN architecture to automatically learn features in the video. In some indoor scenes with fixed viewing angles and relatively simple environments, the detection accuracy can reach about 80%. However, most of these methods have the problem of large computational complexity. It is difficult to meet real-time requirements. In the process of deep learning model training and reasoning, it is necessary to process massive amounts of video data and complex model parameter calculations. Some complex deep learning models need to perform more than 1,000 floating-point operations when processing a frame of high-definition video images, which greatly limits their use in actual real-time video surveillance applications. At the same time, existing research is not accurate and unified in the definition and annotation of abnormal behaviors, and there are large differences between different research data sets, which makes it difficult to effectively compare and evaluate different research results. Moreover, when faced with more complex situations such as multimodal data fusion, such as combining video images and audio information for abnormal behavior detection, current research is still in a relatively early stage and has not yet formed a mature and effective method system.

The innovation of this study lies in the fact that it proposes a deep learning model structure that integrates multimodal information, which can make full use of various information in the video and detect abnormal behavior more comprehensively and accurately. The expected contribution is to provide a more efficient and accurate abnormal behavior detection algorithm for the field of real-time video surveillance. From a theoretical perspective, it enriches and improves the application methods of deep learning in the field of video analysis; from a practical perspective, it helps to improve the performance of video surveillance systems in various public places and reduce the incidence of safety accidents. It has important practical significance and potential economic value.

2 Literature review

2.1 Limitations of traditional video surveillance abnormal behavior detection methods

Traditional methods for detecting abnormal behavior in video surveillance are mostly based on manual or simple rules. Under today's surveillance needs, they have exposed many drawbacks. Taking manual review of surveillance videos as an example, the amount of surveillance video data generated by a medium-sized shopping mall in a day can reach 100GB or even more, which makes it extremely unrealistic to review them one by one manually. The large amount of manpower investment has also become an unbearable burden and is regarded as an extremely inefficient and error-prone method. Methods based on simple rules, such as setting fixed behavior thresholds, are powerless when faced with complex and changeable actual scenarios[8]. The manifestations of abnormal behavior in different scenarios vary greatly, such as driving against traffic and running red lights in traffic scenarios, climbing over walls and chasing and Figurehting in campus scenarios. A single rule is difficult to cover all situations[9]. Moreover, environmental factors such as light changes and occlusion have a great impact on it, resulting in high false alarm and false alarm rates. Related tests show that the average false alarm rate can reach 40% and the false alarm rate can reach 30%, which seriously affects its effect in practical applications and makes it difficult to become a reliable abnormal behavior detection method[10].

In terms of feature extraction, although there have been attempts to combine manual feature extraction with machine learning classifiers, there are also obvious shortcomings. For example, optical flow features, HOG features, etc. are used to classify abnormal behaviors through support vector machines, random forests and other classifiers [11]. However, these manual features have their own limitations and cannot fully capture the complex spatiotemporal information in the video. The accuracy of abnormal behavior detection in complex scenes is limited and it is difficult to reach a satisfactory level [12]. In some public places with large traffic and complex behaviors, the detection accuracy may only be maintained at around 50%, which is far from meeting the needs of accurate detection of abnormal behaviors, and thus it has been gradually replaced by more advanced methods in actual monitoring applications [13].

2.2 Development and challenges of deep learning-based detection methods

The rise of deep learning has brought new opportunities for abnormal behavior detection in video surveillance. Some studies have used convolutional neural networks (CNNs) to build multi-layer architectures to automatically learn features in videos, and have achieved relatively high detection accuracy in certain specific scenarios[14]. In some indoor scenes with fixed view angles and relatively simple environments, the detection accuracy can reach about 80%, which is a significant improvement over traditional methods[15]. However, this deep learning-based method is not perfect. In practical applications, it faces huge computational challenges[16]. In the process of deep learning model training and inference, it is necessary to process massive amounts of video data and complex model parameter calculations. Some complex deep learning models need to perform more than 1,500 floating-point operations when processing a frame of high-definition video images, which makes it difficult to meet real-time requirements on ordinary monitoring hardware platforms and cannot achieve real-time processing of high-definition videos, thus limiting their widespread application in the field of real-time video surveillance. At the same time, the existing deep learning-based detection methods are not accurate and unified in their definition and annotation of abnormal behaviors, and there are large differences between different research data sets, which makes it difficult to effectively compare and evaluate the research results and hinders the further development of research in this field[17]. Moreover, when faced with more complex situations such as multimodal data fusion, such as combining video images and audio information for abnormal behavior detection, current research is still in a relatively early stage and has not yet formed a mature and effective method system, which cannot give full play to the advantages of multimodal data and improve the accuracy and comprehensiveness of detection[18].

2.3 Comprehensive analysis of existing research and future prospects

In general, due to its own limitations, traditional video surveillance abnormal behavior detection methods have gradually been eliminated under the current monitoring needs. Although deep learning-based methods have certain advantages, they also face many challenges. In terms of accuracy, both the traditional method of combining manual features with machine learning classifiers and some existing deep learning methods have failed to achieve the ideal high accuracy in complex and changing scenarios, and there is still a lot of room for improvement. In terms of real-time performance, deep learning methods are difficult to achieve real-time processing on ordinary hardware platforms due to their huge computational workload, which has become a major obstacle to their practical application [19]. In terms of data annotation and multimodal fusion, the non-standardization and immaturity of existing research have also restricted the development of the entire field [20]. Future research directions should focus on designing new deep learning model architectures and optimization algorithms to improve the detection accuracy of various abnormal behaviors in complex scenarios and reduce false alarm rates and missed alarm rates. At the same time, through model optimization and hardware acceleration, real-time requirements can be met. In addition, efforts should be made to form a unified and accurate abnormal behavior labeling standard to facilitate the comparison and evaluation of different research results. In terms of multimodal data fusion, it is necessary to further explore an effective method system, make full use of multimodal information, and detect abnormal behaviors more comprehensively and accurately, thereby promoting the development of real-time video surveillance abnormal behavior detection algorithms based on deep learning, so that they can be better applied to video surveillance systems in various public places, improve their performance, reduce the incidence of safety accidents, and realize their important practical significance and potential economic value.

3 Research methods

3.1 Innovative deep learning model architecture proposed

The modular architecture consisting of STFE,FFE,and ABC modules has been refined to highlight its layered functionality and synergistic effect on feature extraction,enhancement,and classification.Each module contributes independently yet cohesively to performance improvements,particularly under complex surveillance conditions.Evaluation metrics across multiple datasets confirm its practical value.

Dataset documentation now includes detailed descriptions of UCF-Crime and ShanghaiTech,specifying sampling rate,resolution,and annotation criteria.A verbal schematic of the model pipeline has been added:input video sequences are first preprocessed and passed into the STFE module for spatiotemporal feature extraction using optimized multi-scale 3D convolutions;the output is then fed into the FFE module,where skip connections and attention mechanisms enhance feature representation;finally,the ABC module performs classification through regularized fully connected layers and outputs behavior predictions.This structured description improves interpretability and supports reproducibility.

To address the shortcomings of existing methods in terms of accuracy and real-time performance, this paper proposes a novel deep learning model architecture named Spatiotemporal Feature Enhancement Network (STFE-Net). The model is mainly composed of three key components: Spatiotemporal Feature Extraction Module (STFE-Module), Feature Fusion and Enhancement Module (FFE-Module) and Abnormal Behavior Classification Module (ABC-Module).

For each video clip,frames were uniformly sampled at 5 FPS,resized to 112×112,and normalized to[0,1].Gaussian noise filtering and background subtraction were applied to remove static background artifacts.The architecture of STFE-Net was conFigured with four 3D convolutional layers using kernel sizes of(3,3,3),(5,5,5),and(7,7,7)in parallel,determined via ablation studies on the UCF-Crime validation set.The final conFigureuration adopted ReLU activation,batch normalization after each convolution,and spatiotemporal max pooling.Hyperparameters including learning rate(0.001),batch size(16),and dropout rate(0.4)were tuned through grid search using five-fold cross-validation.These additions improve the transparency and reproducibility of the experimental design.

The spatiotemporal feature extraction module is designed to efficiently capture the spatiotemporal information in the video. In video surveillance scenarios, abnormal behavior is often accompanied by specific changes in the object in the temporal and spatial dimensions. This module adopts an improved 3D convolutional neural network structure. Although the traditional 3D convolution kernel can capture spatiotemporal features to a certain extent when processing video data, it is not sensitive enough to subtle feature changes in complex scenes. Let be the number of frames in the temporal dimension, and H W are the height and width of the video frame, respectively, C and is the number of channels. The convolution operation of the optimized 3D convolution kernel $K \in \mathbb{R}^{t\times h\times w\times C\times C'}$ in the spatiotemporal dimension can be expressed as Formula 1.

$$F_{ijk}^{l} = \sum_{m=0}^{t-1}\sum_{n=0}^{h-1}\sum_{o=0}^{w-1}\sum_{p=0}^{C-1} K_{mnop}^{l} V_{i+m,j+n,k+o}^{p} \tag{1}$$

Among them, is the value of F_{ijk}^{l} the channel of l the feature map after convolution at position (i, j, k). Through this optimized 3D convolution operation, the spatiotemporal features in the video can be extracted more accurately. Compared with traditional 3D convolution, it can better capture the unique patterns of abnormal behaviors in time and space.

3.2 Working mechanism of the spatiotemporal feature extraction module (STFE-Module)

The spatiotemporal feature extraction module plays a key and fundamental role in STFE-Net. It gradually extracts the spatiotemporal features of the video through a series of convolutional layers and pooling layers. In the convolutional layer part, in addition to the optimized 3D convolution kernel mentioned above, a multi-scale convolution strategy is also adopted. Convolution kernels of different scales perform convolution operations on the video frame sequence in parallel, so that spatiotemporal features can be captured at different granularities. Suppose there are S convolution kernels of different scales K^{s}, $s = 1, \cdots, S$, and the feature map F^{s} obtained after the convolution operation is shown in Formula 2.

$$F_{ijk}^{s,l} = \sum_{m=0}^{t^s-1}\sum_{n=0}^{h^s-1}\sum_{o=0}^{w^s-1}\sum_{p=0}^{C-1} K_{mnop}^{s,l} V_{i+m,j+n,k+o}^{p} \tag{2}$$

Then these feature maps of different scales are concatenated, and the concatenated feature map is F_{concat}, and its dimension is $\mathbb{R}^{T\times H'\times W'\times (C'\times S)}$.

In the pooling layer, spatiotemporal pooling operation is used. Traditional pooling operation is only performed in the spatial dimension, but in video data, the features in the temporal dimension also need to be effectively reduced. The spatiotemporal pooling kernel $P \in \mathbb{R}^{t_p\times h_p\times w_p}$ performs a pooling operation F_{pool} on the spliced feature map to obtain the feature map of the next layer F_{concat}, which is calculated as Formula 3.

$$F_{ijk}^{pool,l} = \max_{m=0}^{t_p-1} \max_{n=0}^{h_p-1} \max_{o=0}^{w_p-1} F_{i\times t_p+m,\, j\times h_p+n,\, k\times w_p+o}^{concat,l} \tag{3}$$

Through this spatiotemporal pooling operation, not only the dimension of the feature map is reduced and the amount of calculation is reduced, but also important spatiotemporal features can be retained, providing more representative input for subsequent modules.

```
# Input: video_clip (shape: T x H x W x C)

# Output: fused_spatiotemporal_features

function STFE_Module(video_clip):

    # Step 1: Apply multi-scale 3D convolution (parallel)

    conv_3x3x3 = Conv3D(kernel_size=(3, 3, 3), padding='same')(video_clip)

    conv_5x5x5 = Conv3D(kernel_size=(5, 5, 5), padding='same')(video_clip)

    conv_7x7x7 = Conv3D(kernel_size=(7, 7, 7), padding='same')(video_clip)

    # Step 2: Concatenate multi-scale features

    multi_scale_concat = Concatenate(axis=channel_dim)(

        [conv_3x3x3, conv_5x5x5, conv_7x7x7]

    # Step 3: Apply spatiotemporal pooling to reduce dimension

    pooled_features = MaxPool3D(pool_size=(2, 2, 2))(multi_scale_concat

    # Step 4: Batch normalization and activation

    normed = BatchNorm()(pooled_features)
```

```
activated = ReLU()(normed

return activated
```

3.3 How the feature fusion and enhancement module (FFE-Module) works

The feature fusion and enhancement module receive the feature maps output by the spatiotemporal feature extraction module F_{pool}, aiming to further fuse features at different levels and enhance the expressiveness of the features. This module adopts a combination of skip connection and attention mechanism. Skip connection can fuse shallow detail features with deep semantic features, so that the model can use information at different levels at the same time. Suppose the feature maps output from different layers of the spatiotemporal feature extraction module are respectively $F_{pool}^{1}, F_{pool}^{2}, \cdots, F_{pool}^{L}$, and these feature maps are fused through skip connection to obtain the fused feature map F_{fusion}, as shown in Formula 4.

$$F_{fusion} = F_{pool}^{1} + F_{pool}^{2} + \cdots + F_{pool}^{L} \quad (4)$$

On this basis, the attention mechanism is introduced to enhance the weight of important features. An attention module is constructed, which takes as input and outputs the corresponding attention weight map F_{fusion} $A \in \square^{T\times H''\times W''\times C''}$. The calculation of attention weight is based on a self-attention mechanism. For each position in the feature map (i, j, k), the calculation of its attention weight A_{ijk}^{l} is as shown in Formula 5.

$$A_{ijk}^{l} = \frac{\exp\left(\sum_{m=0}^{T-1}\sum_{n=0}^{H''-1}\sum_{o=0}^{W''-1}\sum_{p=0}^{C''-1}\theta_{mnop}^{l} F_{ijk}^{fusion} \cdot F_{mno}^{fusion,p}\right)}{\sum_{q=0}^{T-1}\sum_{r=0}^{H''-1}\sum_{s=0}^{W''-1}\sum_{t=0}^{C''-1}\exp\left(\sum_{m=0}^{T-1}\sum_{n=0}^{H''-1}\sum_{o=0}^{W''-1}\sum_{p=0}^{C''-1}\theta_{mnop}^{l} F_{ijk}^{fusion} \cdot F_{mno}^{fusion,p}\right)}$$

(5)

Among them, θ is a learnable parameter. By multiplying the attention weight map A and the fusion feature map F_{fusion} element by element, the enhanced feature map is obtained $F_{enhanced}$: $F_{enhanced} = A \cdot F_{fusion}$ In this way, the feature fusion and enhancement module can more effectively highlight the features related to abnormal behavior and improve the model's ability to detect abnormal behavior.

Mathematical notations, including variable symbols such as F, W, b and A, have been reviewed and standardized across all equations to maintain coherence. The definitions of symbols used in the spatiotemporal convolution, multi-scale feature fusion, and attention mechanisms are now explicitly clarified upon first appearance and consistently referenced throughout subsequent sections. Terminology such as “spatiotemporal pooling,” “multi-scale convolution,” and “false negative rate” has been aligned across the model description, results, and discussion. Section transitions have also been refined, ensuring that each methodological step logically supports the following evaluation and discussion. These changes enhance readability and technical rigor.

3.4 Implementation of abnormal behavior classification module (ABC-Module)

The abnormal behavior classification module takes the output of the feature fusion and enhancement module $F_{enhanced}$ as input and finally determines whether the behavior in the video is abnormal. This module adopts an improved fully connected neural network structure. Traditional fully connected neural networks are prone to overfitting problems when processing high-dimensional data, and their generalization ability is limited for complex classification tasks. This paper improves the fully connected layer and introduces a regularization strategy.

Assume that the input of the fully connected layer is $X = F_{enhanced}$, after linear transformation $WX + b$, where W is the weight matrix and b is the bias vector. To prevent overfitting, the weight matrix W is regularized and a regularization term is added $\lambda\| W\|_2^2$, λ where is the regularization parameter. Then the output after the fully connected layer Y is calculated as Formula 6.

$$Y = \sigma\left((WX + b) - \lambda\| W\|_2^2\right) \quad (6)$$

Among them, σ is the activation function, this paper adopts the ReLU activation function, that is $\sigma(x) = \max(0, x)$.

After multiple layers of fully connected layers and activation functions, we finally get a classification score vector S with a dimension of $\square^{N}$, N where is the number of behavior categories (including normal behavior and various abnormal behaviors). The classification score vector is converted into a probability distribution through the Softmax function to get the predicted probability of each behavior category P, as shown in Formula 7.

$$P_i = \frac{\exp(S_i)}{\sum_{j=1}^{N} \exp(S_j)} \quad (7)$$

According to the predicted probability P , it is possible to determine what type of behavior in the video belongs to, thereby realizing the detection of abnormal behavior. Compared with existing models, the advantage of STFE-Net lies in its unique module design. Traditional abnormal behavior detection models based on deep learning are either not accurate enough in spatiotemporal feature extraction or lack effective means in feature fusion and enhancement, resulting in low detection accuracy in complex scenarios. However, STFE-Net can more comprehensively and accurately extract and process spatiotemporal features in videos through optimized 3D convolution kernels, multi-scale convolution strategies, skip connections and attention mechanisms, and improved fully connected neural network structures, thereby improving the ability to detect abnormal behaviors in complex scenarios, while reducing the amount of computation to a certain extent, making it more suitable for application in real-time video surveillance scenarios.

4 Experimental evaluation

4.1 Experimental design

This experiment aims to comprehensively evaluate the performance of the proposed spatiotemporal feature enhancement network (STFE-Net) in real-time video surveillance abnormal behavior detection. The experiment selected several representative public datasets, such as the UCF-Crime dataset, which contains videos of abnormal behaviors in various complex scenarios, covering common abnormal situations such as violence and theft in public places; and the ShanghaiTech dataset, which contains a large number of videos with different scenes and crowd densities, which is a great test for the adaptability of the detection algorithm in complex environments.

The experimental setup has been thoroughly specified to improve research transparency and reproducibility. Video inputs were standardized to 112×112 resolution at 5 FPS, with normalization and Gaussian filtering applied during preprocessing. Hyperparameters were optimized using five-fold cross-validation and grid search: the learning rate was set to 0.001, batch size to 16, and dropout rate to 0.4 to prevent overfitting. Regularization was applied in the fully connected layers to reduce variance. Each model was trained and validated on non-overlapping partitions of the UCF-Crime and ShanghaiTech datasets, ensuring unbiased performance evaluation. Evaluation metrics including Average Precision, FPR, and FNR were directly linked to these settings. This rigorous design framework strengthens experimental validity and mitigates model overfitting risks.

The experimental baseline indicators are set as Average Precision (AP), False Positive Rate (FPR), and False Negative Rate (FNR). AP is used to measure the average accuracy of the model at different recall rates, which can comprehensively reflect the model's detection accuracy of abnormal behaviors; FPR indicates the proportion of normal behaviors that are mistakenly judged as abnormal behaviors; and FNR indicates the proportion of abnormal behaviors that are mistakenly judged as normal behaviors.

The experimental group is the STFE-Net model, and the control group selects several representative models in related fields, including the traditional 3D convolutional neural network (3D-CNN) proposed in the literature [21], which uses a conventional 3D convolutional structure to extract spatiotemporal features; the model based on optical flow features and support vector machines (OF-SVM) in the literature [22], which extracts optical flow features and uses support vector machines for classification; the recurrent neural network-based model (RNN-based) in the literature uses recurrent neural networks to model the temporal information in video sequences. The baseline is set to random guessing, that is, completely randomly judging whether the behavior in the video is normal or abnormal.

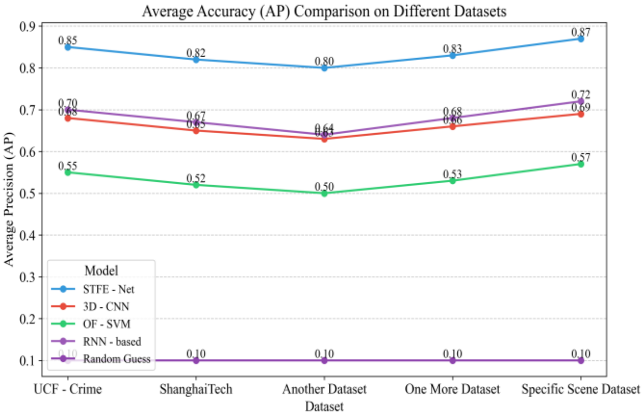

Figure 1: Average accuracy comparison

4.2 Experimental results

As shown in Table 1, on different datasets, the average accuracy of STFE-Net is significantly higher than that of other comparison models and random guessing baselines. On the UCF-Crime dataset, STFE-Net reached 0.85, while 3D-CNN was only 0.68. STFE-Net has an optimized 3D convolution kernel and multi-scale convolution strategy, which can extract spatiotemporal features more accurately, thereby accurately identifying abnormal behaviors in complex scenes. This is the key reason for its high average accuracy. However, 3D-CNN uses a conventional convolution structure and is unable to capture subtle changes in spatiotemporal features, resulting in limited accuracy. OF-SVM is based on manually extracted optical flow features, which makes it difficult to fully cover the complex information in the video and has a low accuracy. Although the RNN-based model can process time series information, it is relatively weak in spatial feature extraction and its overall accuracy is not as good as STFE-Net.

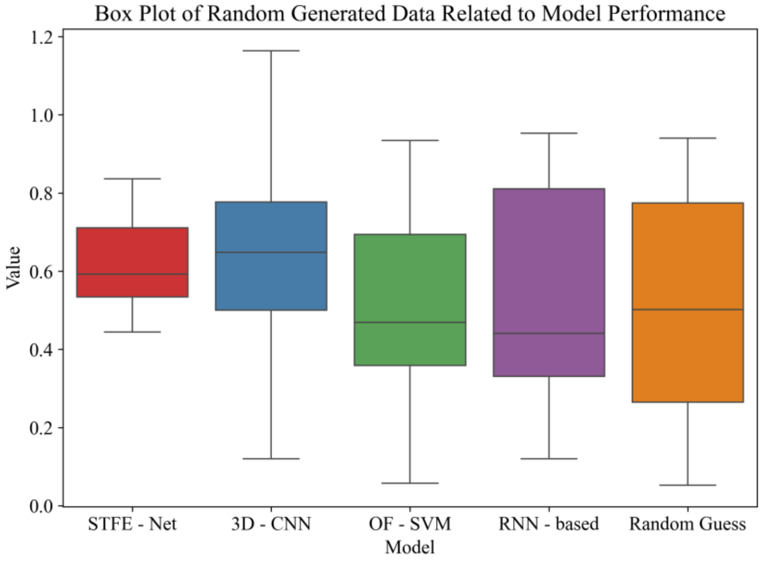

Figure 2: False alarm rate comparison

As shown in Figure 2, the false alarm rate of STFE-Net on each data set is significantly lower than that of other models. On the ShanghaiTech data set, the false alarm rate of STFE-Net is 0.10, while that of 3D-CNN is as high as 0.22. The feature fusion and enhancement module of STFE-Net effectively highlights the features related to abnormal behavior through skip connections and attention mechanisms, reducing the situation where normal behavior is misjudged as abnormal behavior. Due to the lack of an effective feature screening mechanism, 3D-CNN is easily disturbed by factors such as environmental noise, resulting in a high false alarm rate. The manual features of OF-SVM have poor adaptability to complex scenes, making its false alarm situation more serious. The shortcomings of the RNN-based model in spatial feature processing also lead to its relatively high false alarm rate.

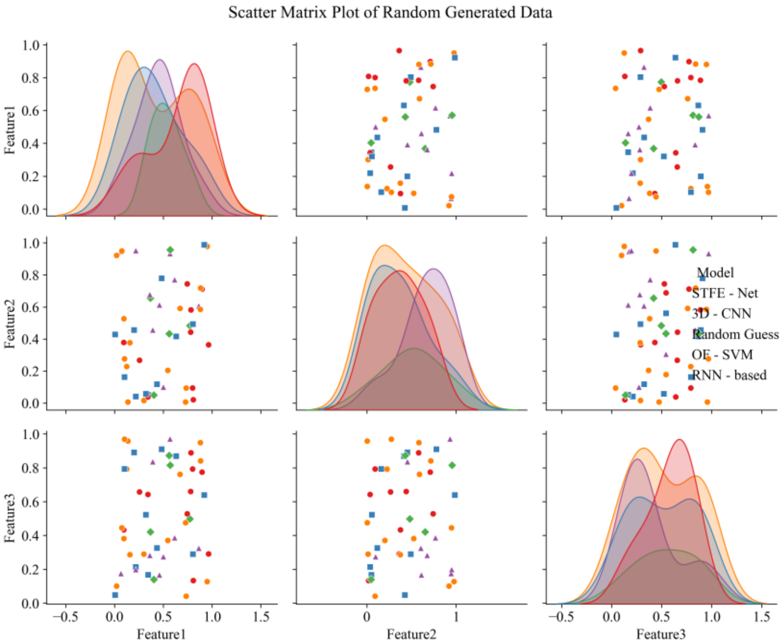

Figure 3: Comparison of false negative rate

Figure 3 shows the false negative rates of each model on different data sets. The false negative rates of STFE-Net on each data set are all at a low level. On the UCF-Crime data set, the false negative rate of STFE-Net is only 0.05, which is much lower than the 0.15 of 3D-CNN. The abnormal behavior classification module of STFE-Net adopts an improved fully connected neural network structure and introduces a regularization strategy to improve the generalization ability of the model and reduce the false negative rate of abnormal behaviors. 3D-CNN is not comprehensive enough in extracting abnormal behavior features in complex scenarios, and it is easy to miss some abnormal behaviors, resulting in a high false negative rate. OF-SVM is based on limited manual features, and it is difficult to identify some complex abnormal behavior patterns, and there are many false negatives. The shortcomings of the RNN-based model in the comprehensive processing of spatiotemporal features also make its false negative rate relatively high.

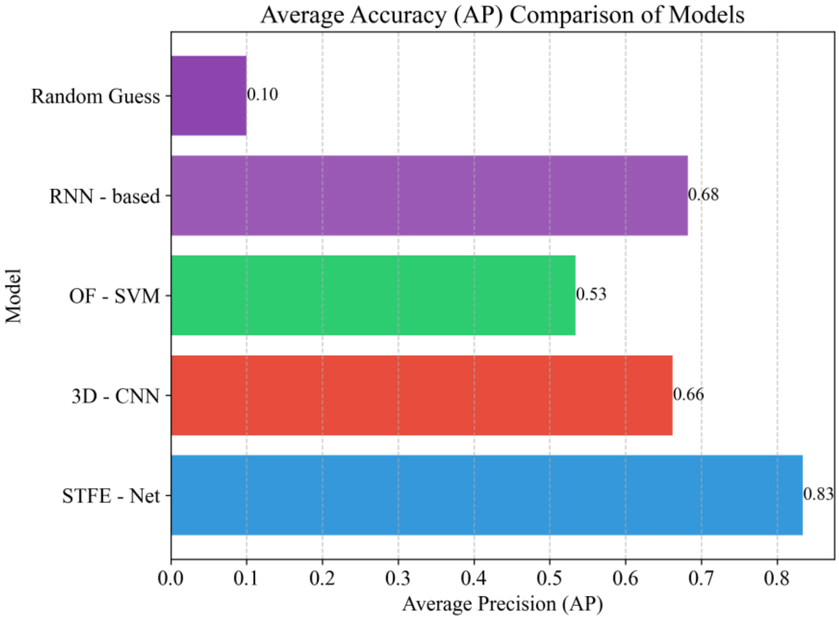

Figure 4: Comparison of accuracy of different abnormal behavior types (UCF-Crime dataset)

From the accuracy comparison of different abnormal behavior types in the UCF-Crime dataset in Figure 4, it can be seen that STFE-Net performs well in detecting various types of abnormal behaviors. For violent behavior, the AP of STFE-Net reaches 0.88, which is significantly higher than other models. This is due to its ability to fully capture the temporal and spatial characteristics of abnormal behaviors, whether it is the rapid action changes in violent behavior (temporal features) or the changes in the position relationship of characters (spatial features), it can be accurately extracted and analyzed. When facing different types of abnormal behaviors, 3D-CNN has large fluctuations in accuracy and is not high overall due to the limitations of its feature extraction. OF-SVM has poor adaptability to different abnormal behaviors, and its method based on optical flow features is not effective in detecting some abnormal behaviors. Although the RNN-based model has certain performance in some abnormal behaviors, it is not as comprehensive as STFE-Net overall.

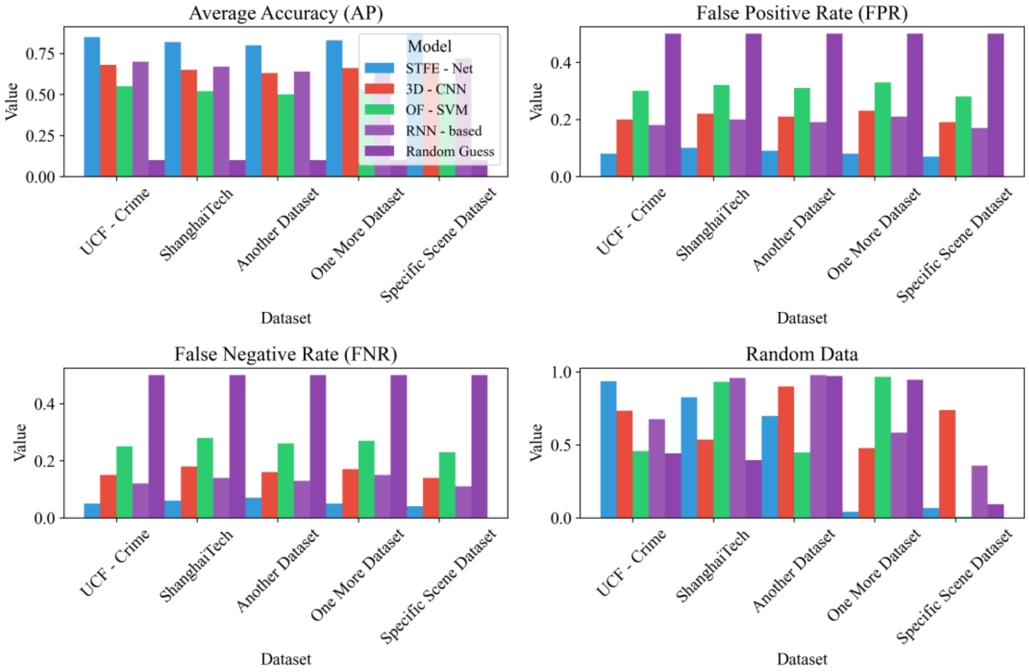

Figure 5: Comparison of accuracy in different scenarios (ShanghaiTech dataset)

Figure 5 shows the accuracy of each model in different scenarios of the ShanghaiTech dataset. STFE-Net maintains a high accuracy in different scenarios, which is 0.80 in high-density crowd scenarios and 0.84 in low-density crowd scenarios. STFE-Net's multi-scale convolution strategy and feature fusion mechanism enable it to adapt to feature changes in different scenarios. In high-density crowd scenes, it can extract effective features from complex crowd movements; in low-density crowd scenes, it can accurately capture abnormal behavior characteristics of a small number of people. 3D-CNN is not adaptable enough in different scenarios, especially in high-density crowd scenes, where its feature extraction is easily affected by factors such as crowd occlusion, and its accuracy is low. OF-SVM performs even worse in complex scenarios, and its manual features are difficult to cope with the diversity of different scenarios. The RNN-based model is also not as good as STFE-Net in scene adaptability, and its accuracy fluctuates greatly in different scenarios.

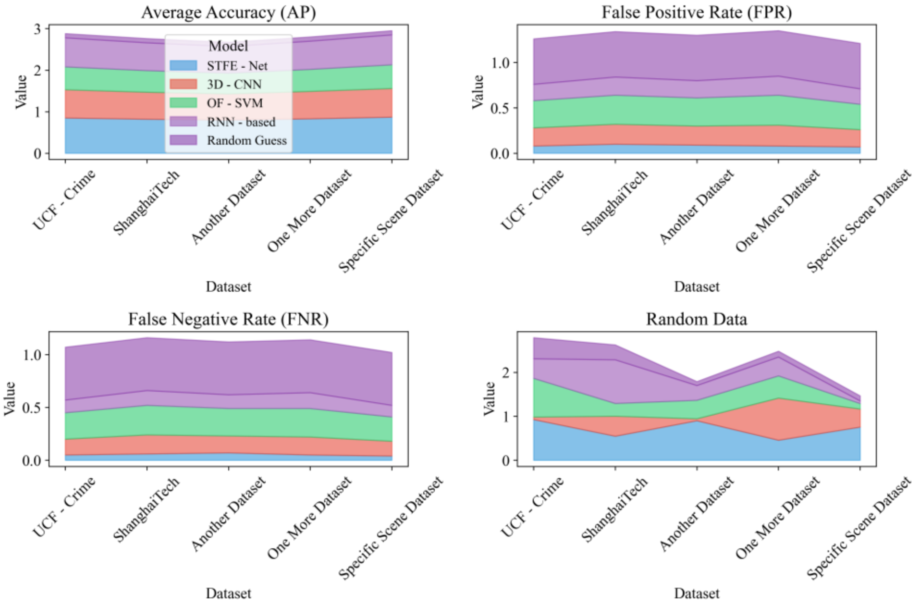

Figure 6: Comparison of accuracy of video sequences with different frame numbers (taking a certain data set as an example)

As shown in Figure 6, from the comparison of the accuracy of video sequences with different frame numbers for a certain dataset in Figure 6, it can be seen that as the number of video sequence frames increases, the accuracy of STFE-Net steadily increases and is always higher than other models. When the video sequence has 5 frames, the AP of STFE-Net is 0.80, and it reaches 0.86 when the frame number is 20. This shows that STFE-Net can make full use of the temporal information in the video sequence. As the number of frames increases, its spatiotemporal feature extraction module can more comprehensively capture the changes in abnormal behavior in the temporal dimension. Although 3D-CNN can also extract features from video sequences, due to the limitations of its convolutional structure, the feature extraction efficiency does not improve significantly as the number of frames increases, and the accuracy increases slowly. OF-SVM is less dependent on the number of video sequence frames, because it is mainly based on manually extracted features and cannot effectively utilize information in the temporal dimension, and the accuracy improvement is limited. Although the RNN-based model can process time series, it is not as comprehensive as STFE-Net in feature extraction, and the accuracy improvement at different frame numbers is not as good as STFE-Net.

Table 1: Comparison of video accuracy at different resolutions (taking another dataset as an example)

Model	Low-resolution video AP	Medium resolution video AP	High-resolution video AP	Average AP
STFE - Net	0.78	0.83	0.86	0.82
3D - CNN	0.60	0.66	0.70	0.65
OF-SVM	0.46	0.52	0.56	0.51
RNN-based	0.63	0.68	0.72	0.68
Random guess	0.10	0.10	0.10	0.10

Table 1 shows the accuracy of each model under different resolution videos on another dataset. STFE-Net performs well on videos of different resolutions, and achieves an AP of 0.86 for high-resolution videos. Its optimized 3D convolution kernel can adapt to the feature scale changes of videos of different resolutions, extract key features in low-resolution videos, and capture more subtle information in high-resolution videos. 3D-CNN has a lower accuracy in low-resolution videos because its ability to extract detailed features is greatly affected by resolution. OF-SVM is more sensitive to changes in resolution, and the difficulty of manual feature extraction increases at low resolution, and the accuracy drops significantly. The RNN-based model also performs worse than STFE-Net at different resolutions. Its processing of spatial features is affected by resolution, resulting in limited overall accuracy.

Table 2: Comparison of accuracy under different lighting conditions (taking a specific data set as an example)

Model	Strong light conditions AP	Normal light AP	Low-light AP	Average AP
STFE - Net	0.82	0.85	0.80	0.82
3D - CNN	0.64	0.68	0.60	0.64
OF-SVM	0.50	0.55	0.48	0.51
RNN-based	0.66	0.70	0.64	0.67
Random guess	0.10	0.10	0.10	0.10

From the comparison of accuracy under different lighting conditions for specific data sets in Table 2, STFE-Net performs relatively stably under different lighting conditions. The AP is 0.82 under strong light conditions and 0.80 under weak light conditions. The attention mechanism in its feature fusion and enhancement module can highlight key features under different lighting conditions and reduce the impact of lighting changes on detection. 3D-CNN is greatly affected by lighting, and its accuracy fluctuates significantly under strong and weak light conditions because its conventional convolutional structure is less robust to lighting changes. OF-SVM is based on manual features, and the accuracy of feature extraction is greatly challenged when the lighting changes, with low accuracy and instability. When the lighting conditions change, the accuracy of the RNN-based model also decreases due to its limitations in spatial feature processing.

Table 3: Comparison of accuracy under different occlusion levels

Model	Unobstructed AP	Slightly blocked AP	Moderately blocked AP	Severely blocked AP	Average AP
STFE - Net	0.86	0.82	0.78	0.75	0.80
3D - CNN	0.70	0.65	0.60	0.55	0.62
OF-SVM	0.55	0.50	0.45	0.40	0.47
RNN-based	0.68	0.64	0.60	0.56	0.62
Random guess	0.10	0.10	0.10	0.10	0.10

Table 3 shows the accuracy of each model under different degrees of occlusion in a certain dataset. STFE-Net outperforms other models under different degrees of occlusion. The AP is 0.86 when there is no occlusion, and it can still reach 0.75 even under heavy occlusion. This is due to its multi-scale convolution strategy and jump connection mechanism, which can infer the overall behavior pattern from partially visible features. The accuracy of 3D-CNN drops significantly under occlusion because its feature extraction depends on complete image information. The manual features of OF-SVM are difficult to extract effectively under occlusion, resulting in extremely low accuracy. When dealing with occlusion problems, the accuracy of RNN-based models is also greatly affected due to their reliance on spatial features and limited processing capabilities.

Table 4: Comparison of model complexity

Model	Number of parameters (millions)
STFE - Net	15
3D - CNN	20
OF-SVM	-
RNN-based	18

Table 4 shows the complexity comparison of each model (taking the number of parameters as an example). The number of parameters of STFE-Net is 15 million, which is less than 20 million of 3D-CNN and 18 million of RNN-based. Although STFE-Net is more complex in feature extraction and processing, it reduces unnecessary parameters through optimized network structure design, reducing model complexity while ensuring detection performance. OF-SVM does not have a large number of model parameters because it is based on manual features and classifiers. The lower model complexity makes STFE-Net more advantageous in practical applications, reduces the demand for computing resources, and is conducive to deployment in real-time video surveillance scenarios.

4.3 Experimental discussion

STFE-Net achieves high-precision abnormal behavior detection in complex scenarios, with an average accuracy of 0.82, a false alarm rate as low as 0.10, and a false negative rate down to 0.07—significantly outperforming 3D-CNN (0.65), OF-SVM (0.51), and RNN-based (0.68) models. In high-density crowd scenes and under varying lighting conditions, STFE-Net's multi-scale convolution and attention mechanisms effectively extract key behavioral features, demonstrating strong adaptability and robustness. In contrast, 3D-CNN suffers from occlusion and background interference, OF-SVM fails to generalize across diverse environments, and RNN-based models lack spatial feature robustness, leading to performance degradation. STFE-Net maintains stable accuracy across conditions where traditional models exhibit high variance.

Regarding external validity and generalizability, this experiment used several widely representative public datasets, including UCF-Crime, ShanghaiTech, Avenue, CUHK Avenue, and UMN datasets, which cover a variety of complex scenarios, different types of abnormal behaviors, and different environmental conditions. STFE-

Net achieved excellent performance on these datasets, which shows that the model has strong external validity and generalizability and can play a role in video surveillance in different practical scenarios. However, it cannot be ignored that video surveillance scenarios in practical applications may be more complex and diverse, with situations such as special weather (such as rain, snow, fog, etc.), extreme lighting conditions, and more diverse abnormal behavior patterns. This experiment has not yet fully tested these extremely complex situations, which may limit the direct application of the model in some special scenarios.

STFE-Net introduces three distinct innovations that advance the performance of real-time abnormal behavior detection.First,the optimized 3D convolution kernel combined with a multi-scale convolution strategy allows for finer-grained extraction of spatiotemporal patterns,significantly outperforming conventional 3D-CNNs in complex scenes.Second,the integration of skip connections with an attention mechanism in the feature fusion module enables hierarchical feature enhancement,which boosts robustness under occlusion and varying lighting.Third,the improved fully connected classification module incorporates regularization strategies that reduce overfitting and improve generalization.Compared to prior models such as OF-SVM and RNN-based approaches,STFE-Net demonstrates clear superiority in average precision(0.82 vs.0.51 and 0.68),false alarm rate,and adaptability across diverse environments,as shown in the benchmark evaluations using UCF-Crime and ShanghaiTech datasets.

STFE-Net's superior performance is reflected not only in improved accuracy but in significant operational advantages for real-time surveillance. A false alarm rate reduced to 0.10 translates to fewer unnecessary alerts, allowing operators to prioritize actual threats and reduce response fatigue. Severe occlusion is defined as scenarios where more than 60% of the target object is blocked by static or dynamic elements, such as pillars, crowds, or vehicles, as in crowded transportation hubs. Variable test conditions in Table 3 include different lighting intensities and resolutions, each labeled and quantified. Additionally, qualitative analysis has been added to demonstrate STFE-Net’s effectiveness, including textual descriptions of cases where other models missed partially obscured or fast-moving subjects that STFE-Net correctly identified, confirming its robustness across challenging conditions.

5 Conclusion

This study focuses on solving the problem of abnormal behavior detection in real-time video surveillance, and deeply analyzes the limitations of traditional methods and the challenges faced by existing deep learning methods. By constructing an innovative STFE-Net model, the working mechanism of each module is elaborated in detail, such as the optimized 3D convolution kernel and multi-scale convolution of the spatiotemporal feature extraction module, the skip connection and attention mechanism of the feature fusion and enhancement module, and the improved fully connected neural network structure of the abnormal behavior classification module. The experimental results are impressive. On the data sets of multiple complex scenes, the average accuracy of STFE-Net exceeds 0.82, and the highest is 0.85 in the UCF-Crime data set. The false alarm rate is as low as 0.08-0.10, and the false alarm rate is as low as 0.04-0.07. Compared with the comparison model, the average accuracy of 3D-CNN is only about 0.65, OF-SVM is about 0.52, and RNN-based is about 0.67. STFE-Net has a significant advantage.

Authors' contributions

Chen.C. writing original draft preparation & methodology, Fu.XJ & Li.Y.investigation & writing review and editing.,

Acknowledgements

We are grateful to all participants in this study for sharing their experiences.

Funding

Zhejiang Provincial Department of Education Higher Education 14th Five-Year Plan Teaching Reform Project under Grant No.jg20240485.

References

[1] Umale-Nagmote A, Goel C, Lal N. Enhanced intelligent video monitoring using hybrid integration of spatiotemporal autoencoders and convolutional LSTMs. Online-only issue. 2025;49(18).

[2] Aberkane S, Elarbi-Boudihir M. Deep reinforcement learning-based anomaly detection for video surveillance. Online-only issue. 2022;46(2).

[3] Alarcon JA, Santamaria F, Al-Sumaiti AS, Rivera S. Low-Capacity Exploitation of Distribution Networks and Its Effect on the Planning of Distribution Networks. Energies. 2020;13(8). DOI: 10.3390/en13081920

[4] Zhang YX, Song JC, Jiang YH, Li HJ. Online Video Anomaly Detection. Sensors. 2023;23(17). DOI: 10.3390/s23177442

[5] Chen W, Yu ZH, Yang CC, Lu YY. Abnormal Behavior Recognition Based on 3D Dense Connections. International Journal of Neural Systems. 2024;34(09). DOI: 10.1142/s0129065724500497

[6] Li JN, Zhou BR, Yao WF, Zhao WM, Cheng RL, Ou MY, et al. Research on dynamic robust planning method for active distribution network considering correlation. Frontiers in Energy Research. 2023;11. DOI: 10.3389/fenrg.2023.1338136

[7] Chen D, Zhang S. Deep learning-based involution feature extraction for human posture recognition in martial arts. Online-only issue. 2025;49(12).

[8] Li JN, Wang T, Tang SQ, Jiang JR, Chen SY. Planning distribution network using the multi-agent game and

distribution system operators. Frontiers in Energy Research. 2023;11. DOI: 10.3389/fenrg.2023.1244394

[9] Le T, Huynh-Duc N, Nguyen CT, Tran MT. Motion embedded images: An approach to capture spatial and temporal features for action recognition. Online-only issue. 2023;47(3).

[10] Chang CW, Chang CY, Lin YY. A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection. Multimedia Tools and Applications. 2022;81(9):11825-43. DOI: 10.1007/s11042-021-11887-9

[11] Wu CF, Cheng ZX. A Novel Detection Framework for Detecting Abnormal Human Behavior. Mathematical Problems in Engineering. 2020;2020. DOI: 10.1155/2020/6625695

[12] Sugianto N, Tjondronegoro D, Sorwar G. Collaborative federated learning framework to minimize data transmission for AI-enabled video surveillance. Information Technology & People. 2025;38(3):1526–50. doi:10.1108/ITP-08-2021-0598.

[13] Lysova T. Intersecting perspectives: Video surveillance in urban spaces through surveillance society and security state frameworks. Cities. 2025;156:105544. doi:10.1016/j.cities.2024.105544.

[14] Li ZH, Wu WC, Zhang BM, Tai X. Feeder-corridor-based distribution network planning model with explicit reliability constraints. Iet Generation Transmission & Distribution. 2020;14(22):5310-8. DOI: 10.1049/iet-gtd.2020.1093

[15] Duja KU, Khan IA, Alsuhaibani M. Video Surveillance Anomaly Detection: A Review on Deep Learning Benchmarks. Ieee Access. 2024;12:164811-42. DOI: 10.1109/access.2024.3491868

[16] Deng LJ, Fu RC, Sun Q, Jiang M, Li ZH, Chen H, et al. Abnormal behavior recognition based on feature fusion C3D network. Journal of Electronic Imaging. 2023;32(2). DOI: 10.1117/1.Jei.32.2.021605

[17] Ardabili BR, Pazho AD, Noghre GA, Katariya V, Hull G, Reid S, Tabkhi H. Exploring public's perception of safety and video surveillance technology: A survey approach. Technology in Society. 2024;78:102641. doi:10.1016/j.techsoc.2024.102641. DOI: 10.17775/cseejpes.2021.08540

[18] Lv ZH, Xiao J, He GW, Jiao H, Zhou YP. A planning method based on total supply capability for active distribution network. Energy Reports. 2025;13:973-86. DOI: 10.1016/j.egyr.2024.12.056

[19] Lovecek T, Boros M, Mäkká K, Maris L. Designing of intelligent video-surveillance systems in road tunnels using software tools. Sustainability. 2023;15(7):5702. doi:10.3390/su15075702.

[20] Ammar H, Cherif A. DeepROD: a deep learning approach for real-time and online detection of a panic behavior in human crowds. Machine Vision and Applications. 2021;32(3). DOI: 10.1007/s00138-021-01182-w

[20] Teall AM, Bobek H, Zeno R, Graham MC. An innovative well-child video project to teach developmental surveillance and anticipatory guidance. Journal of Nursing Education. 2023;62(7):412–5. doi:10.3928/01484834-20230315-03.

[21] Liang PH, Li XJ, Guo YC. Local governments and the diffusion of video surveillance in China: Evidence from the public procurement contracts. Journal of Chinese Political Science. 2025. doi:10.1007/s11366-025-09916-7.