

Article

A Model for Detecting Abnormal Elevator Passenger Behavior Based on Video Classification

Jingsheng Lei ^{1,2}, Wanfa Sun ¹, Yuhao Fang ³, Ning Ye ¹, Shengying Yang ¹  and Jianfeng Wu ^{4,*}¹ School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China; 222208855023@zust.edu.cn (W.S.); syyang@zust.edu.cn (S.Y.)² Zhejiang Xinzailing Technology Co., Ltd., Hangzhou 310051, China³ School of Computer and Information Technology, Hefei University of Technology, Xuancheng 242000, China⁴ School of Information Science and Technology, Zhejiang Shuren University, Hangzhou 310015, China

* Correspondence: wujianfeng@zjsru.edu.cn

Abstract: In the task of human behavior detection, video classification based on deep learning has become a prevalent technique. The existing models are limited due to an inadequate understanding of behavior characteristics, which restricts their ability to achieve more accurate recognition results. To address this issue, this paper proposes a new model, which is an improvement upon the existing PPTSM model. Specifically, our model employs a multi-scale dilated attention mechanism, which enables the model to integrate multi-scale semantic information and capture characteristic information of abnormal human behavior more effectively. Additionally, to enhance the characteristic information of human behavior, we propose a gradient flow feature information fusion module that integrates high-level semantic features with low-level detail features, enabling the network to extract more comprehensive features. Experiments conducted on an elevator passenger dataset containing four abnormal behaviors (door picking, jumping, kicking, and door blocking) show that the top-1 Acc of our model is improved by 10% compared to the PPTSM model, reaching 95%. Moreover, experiments with four publicly available datasets (UCF24, UCF101, HMDB51, and the Something-Something-v1 dataset) demonstrate that our method achieves results superior to PPTSM by 6.8%, 6.1%, 21.2%, and 3.96%, respectively.

Keywords: abnormal behavior detection; attention mechanism; elevator accident; temporal shift module



Citation: Lei, J.; Sun, W.; Fang, Y.; Ye, N.; Yang, S.; Wu, J. A Model for Detecting Abnormal Elevator Passenger Behavior Based on Video Classification. *Electronics* **2024**, *13*, 2472. <https://doi.org/10.3390/electronics13132472>

Academic Editor: Ping-Feng Pai

Received: 20 April 2024

Revised: 12 June 2024

Accepted: 18 June 2024

Published: 24 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Elevators serve as essential vertical conveyance mechanisms within modern urban infrastructure, providing convenience for people in daily work and life, with safety as a top priority [1,2]. Incidences of irregular conduct within these confined spaces—ranging from door picking, jumping, kicking, door blocking, inadvertent falls, or misplaced items to deliberate acts of vandalism—can precipitate safety risks. Such behaviors not only endanger the well-being of passengers but also compromise the integrity of the elevator system [3–5].

In recent years, machine learning methodologies have been employed for the detection of aberrant behaviors in elevators. Zhu et al. [6] extracted people and objects from surveillance videos through background subtraction, counted the number of people in a picture and combined the image entropy of motion history images (MHIs) to determine whether passengers fell or acted violently. Sun et al. [7] introduced a detection approach based on the kinetic energy of corners to identify instances of aggression among elevator occupants. In a similar vein, Liu et al. [8] harnessed multi-feature fusion, combined with machine vision, to detect falls of passengers within the elevator confines. While these strategies have shown a capability to identify unusual human behaviors, they are beset with challenges such as inadequate feature extraction capabilities from images and videos and the failure to effectively model temporal information.

The advent of deep learning has heralded remarkable advancements in the realms of video comprehension and human activity recognition [9]. Lan et al. [10] put forth a

dual-stream neural network algorithm for the detection of anomalous behaviors, such as falls, physical altercations, and tampering with doors within elevator environments. Chen, Y., et al. [11] developed an enhanced two-stream neural network architecture, leveraging a 3D ResNet framework, specifically tailored for the detection of falls among elevator passengers through the application of edge computing technology. Shi, Y., et al. [12] used the key point detection algorithm Openpose to detect human skeleton points and detect abnormal door blocking and door-picking behaviors among passengers.

While the previously mentioned methods are capable of identifying abnormal behaviors of passengers in elevator cabins, they encounter challenges such as substantial computational demand, sophisticated detection algorithms, and stringent hardware requirements.

This paper utilizes the PPTSM network as the baseline model for implementing the detection of elevator passengers' abnormal behavior and proposes improvements to the PPTSM network. To detect the abnormal behavior of passengers in elevators, this paper proposes a video classification model named Temporal Shift Module Network with Col-Depth-Point Convolution and Multi-Scale Dilated Attention (TSM-CDPMSDANet). This model leverages the capabilities of the ResNet50 backbone network, combined with the Temporal Shift Module (TSM) [13].

Initially, we embarked on gathering a diverse array of passenger behavior videos within elevator cabins, focusing on four primary behaviors: door picking, jumping, kicking, and door blocking. This effort led to the creation of a comprehensive dataset dedicated to passenger abnormal behavior. Subsequently, we enhanced the ResNet50 architecture by integrating it with the TSM, which was then incorporated into the forefront of a residual block in the network. Notably, the inclusion of a Multi-Scale Dilated Attention (MSDA) [14] module enables our network to concentrate on nuanced features indicative of abnormal behaviors, significantly boosting its capability to detect such activities with heightened accuracy. Further advancements were made through the development of a gradient flow feature fusion module. This innovative component merges high-level semantic features with intricate low-level details, effectively minimizing the network's parameter count while simultaneously augmenting its recognition capabilities. In the final stage, we implemented a data augmentation strategy throughout the training process to enrich dataset diversity and bolster the model's robustness. This was harmoniously combined with the utilization of the cross-entropy loss function to effectively address and mitigate issues arising from imbalances in data categories. The empirical evidence from our experiments convincingly demonstrates that our model substantially elevates the detection accuracy of the four highlighted abnormal behaviors, marking a significant improvement in its diagnostic prowess.

2. Related Work

2.1. Human Behavior Recognition

Traditional human detection techniques rely on manually created features and template matching methods. Gal et al. [15] proposed a human detection method that combines random forest and Hough transform. Although this method shows good results under specific conditions, this method is limited due to feature selection and extraction, and it cannot capture the complex context and diversity features required for human detection. The development of deep learning has brought new possibilities to human body detection. Donahue et al. [16] used a single-stream method to successfully integrate the two deep learning models of a convolutional neural network and a long short-term memory network, making full use of CNN's efficient ability for image feature extraction and LSTM's ability to process the powerful performance in time series data, enabling a more accurate recognition of human movements. Simonyan [17] and Feichtenhofer, C., et al. [18] adopted the dual-flow method to achieve the effective recognition of behaviors by combining prior information (such as optical flow information) with image features. TS-LSTM and temporal inception [19] are based on LSTM, and they fuse high-level spatial and temporal features to learn hidden features over time. By properly leveraging temporal information at multiple

scales, better performance can be achieved even when feature vectors are used as inputs (i.e., feature maps are not used). However, these methods are based on using only 2D convolution and 2D pooling operations. Using only a 2D CNN network to extract features will have the disadvantage of insufficient extraction of global features via the network.

Three-dimensional convolutional neural networks (3D ConvNets) can comprehensively consider temporal and spatial information, so they show great potential in processing tasks such as video sequences, action behaviors, and motion trajectories, and they are, thus, applied in the field of human action recognition [20–23]. SlowFast [24] representatives believe that the process of human behavior is very fast. Therefore, they conduct research on the special dimension of time and design a specific branch to fuse information with different time resolutions to achieve the effect of capturing high-frame-rate action information. These methods reduce the number of parameters, thereby reducing the hardware resources required to run the model. However, their recognition accuracy of human behavior is not high.

Therefore, the method of using skeletal key points for behavior recognition significantly improves the accuracy of human behavior recognition. The ST-GCN [25] network extracts skeleton sequence features through spatio-temporal graph convolution. Skeleton-based information can capture motion information very well, which greatly improves the network's recognition of human behavior. In view of the unreasonable adjacency matrix strategy used in this network, the two-stream adaptive graph convolution network (2s-AGCN) [26] proposed an improved adjacency matrix strategy. In this network, the topology of the graph is determined via BP. The algorithm is learned end-to-end uniformly or individually to increase the flexibility of the model construction graph and achieve greater versatility in adapting to various data samples, making it more suitable for action recognition tasks. LDT-NET [27] proposes a lightweight behavior recognition model, which is based on skeleton points and designed using a multi-stream neural network architecture. This model uses a multi-stream, deep, separable convolutional neural network to successfully achieve high-efficiency action recognition by integrating feature information at three different scales. Although these methods can improve the accuracy of behavior recognition, they face problems such as high data quality requirements and complex calculations.

2.2. Detection of Abnormal Human Behavior in Elevators

Feng et al. [28] designed an elevator monitoring system by combining infrared sensors and smoke alarms to effectively monitor the abnormal behavior of passengers in an elevator cabin. Zhao et al. [29] detected elevator doors through background subtraction and used the YOLOv3 algorithm to count elevator passengers. However, they did not consider the interference of moving objects during the elevator door detection process and the problem of double counting of persons. Wu et al. [30] implemented the functions of judging the status of elevator doors, counting the non-duplicate flow of elevator passengers, and identifying passenger attributes. Yan Qi et al. [31] designed and developed a video monitoring system for elevator abnormal behavior using the edge computing paradigm. The identification of abnormal image sequences and the assessment of abnormal passenger behavior were realized. Shu et al. [32] used the Lucas–Kanade optical flow method to extract motion speed data and constructed a comprehensive feature vector based on this. This feature vector not only fuses angular momentum information but also summarizes key features of the target, which makes it effective in detecting fights and violent behaviors in elevators. GraftNet [33] launched an advanced solution for fine-grained multi-label recognition tasks, specifically used to accurately identify human body features and perform the anomaly monitoring of human flow data through unsupervised learning, effectively capturing abnormal behaviors that may lead to security risks or illegal activities. TSTR-ResNet [34] introduces a new abnormal elevator passenger behavior recognition technology (TSTR module) based on time offset and reinforcement. This count effectively captures spatiotemporal information through the network module, and it uses wavelet decomposition to extract low-frequency subbands of the image to reduce high-frequency noise interference, thereby improving

the accuracy of abnormal behavior detection. The above methods face the problem of low accuracy in identifying the abnormal behavior of passengers in an elevator.

3. Methods

3.1. Overview

The architecture of our model is illustrated in Figure 1, and it is named the Temporal Shift Module Network with Col-Depth-Point Convolution and Multi-Scale Dilated Attention (TSM-CDPMSDANet), featuring a dual-component design: a backbone and a head. The PPTSM model serves as our baseline implemented via the PaddlePaddle framework, incorporating ResNet50_vd [35] as its backbone network, and it is enriched with the temporal shift module from the TSM model. As depicted in Figure 2, this temporal shift module is seamlessly integrated into the two-dimensional convolutional neural network, thus elevating the network's ability to model temporal information without the necessity of extra computational resources. From Figure 1, it can be seen that the input image extracts deep-level features through the backbone and then performs classification tasks through the head. The MSDA module we proposed is employed before the head because the image can obtain a deep-level feature representation after being processed via the backbone. The CDPWC module we proposed is distributed in each feature extraction stage of the backbone, which can effectively improve model performance.

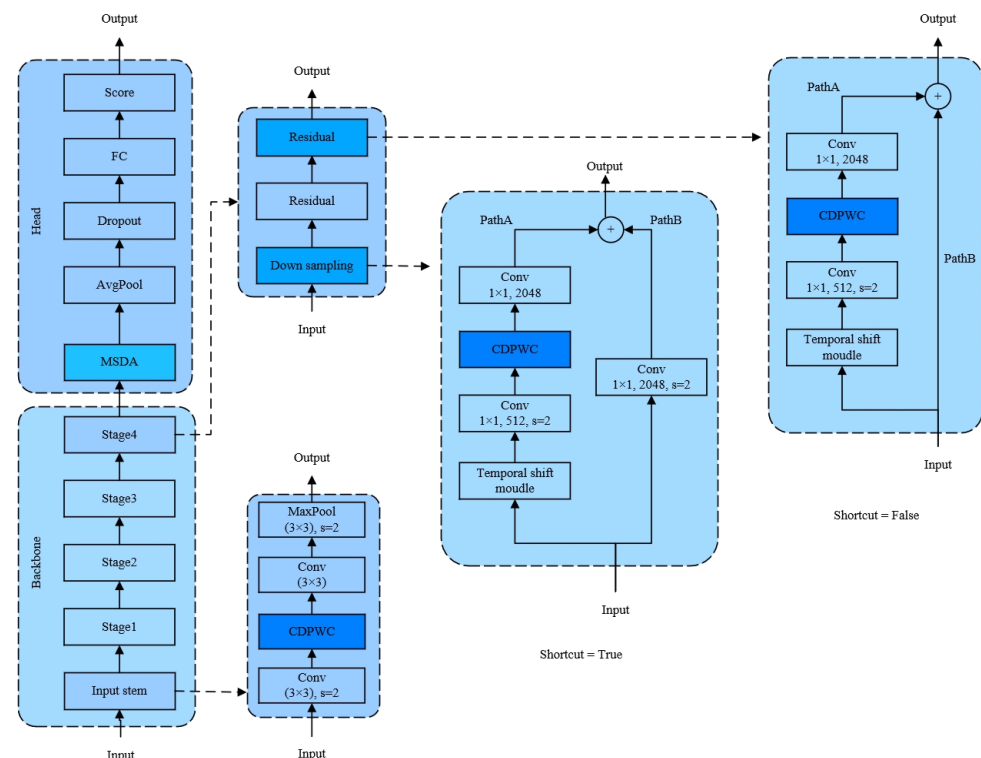


Figure 1. Architecture diagram of Temporal Shift Module Network with Col-Depth-Point Convolution and Multi-Scale Dilated Attention (TSM-CDPMSDANet).

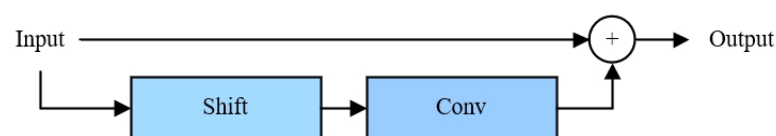


Figure 2. Temporal Shift Module (TSM) with residual structure.

The TSM-CDPMSDANet model initiates its process with the input image undergoing an initial processing block designed to distill preliminary shallow image features. These features are then propelled through four progressive feature extraction stages to unearth

more profound characteristics. At the head of the model, we integrate an MSDA mechanism that orchestrates the interplay between local and sparsely distributed image patches. This facilitates the capture of multi-scale semantic information, enhancing the model's precision in differentiating nuanced disparities across various abnormal behaviors. Concurrently, this enhancement bolsters the model's prowess in handling complex data and elevates its classification accuracy for diverse abnormal activities.

To augment the network's capacity for capturing a richer array of behavioral feature information, our backbone architecture incorporates the sophisticated Col-Depth-Point Wise Convolution(CDPWC) module. This module, which leverages depth-separable convolution, is meticulously integrated at the input-processing level, and it extends through each subsequent feature extraction stage. The deployment of this module ensures optimized gradient flow information, thereby enhancing the fidelity of feature representation and enabling a more nuanced understanding of behavioral patterns. This strategic module insertion trims the overall network's parameter count while simultaneously empowering the model to attain heightened behavioral recognition accuracy. The detailed procedures of the TSM-CDPMSDANet training pipeline are summarized in Algorithm 1.

Algorithm 1 TSM-CDPMSDANet Training Pipeline.

Input: $D(x, y)$: the labeled data x means frames of videos; y means the label.

Output: Trained TSM-CDPMSDANet model $F(\cdot)$.

```

1: Initialize TSM-CDPMSDANet.
2: for iter = 1 to max_iter do
3:   Feed with  $D$ 
4:    $F_0 \leftarrow \text{Input stem}(D)$ ,  $F_0$ : features
5:   for  $i = 0$  to 4 do
6:      $F_{i+1} \leftarrow \text{stage}_{i+1}(F_i)$ 
7:      $F_{bac} \leftarrow F_{i+1}$ 
8:   end for
9:    $F_{AvgPool} \leftarrow \text{AvgPool}(F_{bac})$ 
10:   $F_{Dropout} \leftarrow \text{Dropout}(F_{AvgPool})$ 
11:   $F_{FC} \leftarrow \text{FC}(F_{Dropout})$ 
12:   $Score \leftarrow \text{Score}(F_{FC})$ 
13:   $Output \leftarrow \text{Tensor}(Score, Bbox, Class)$ 
14:   $Detector\ weight \leftarrow Output$ 
15:  Compute Loss.
16:  Optimize TSM-CDPMSDANet by minimizing Loss.
17: end for
```

3.2. Multi-Scale Dilated Attention Module (MSDA)

The primary challenge in detecting passengers' abnormal behaviors within elevator cabins lies in the complexity of fully capturing the diverse manifestations of such behaviors. To tackle this, we introduced MSDA into the head of our model, enabling it to more effectively concentrate on the key features indicative of various behaviors. As depicted in Figure 3, the MSDA module strategically segments the channels of the input feature map into distinct head subsets by employing a multi-head structure (set to 8 in our study). Each head subset undergoes Sliding Window Dilated Attention (SWDA), which assigns unique dilation coefficients to each head, treats a single, original pixel as a token, and selects 8 neighboring pixels based on the dilation coefficient. In our model, a convolution kernel with an expansion rate of 1 can capture local details, while a convolution kernel with an expansion rate of 2 can capture a wider range of global information. This process involves computing self-attention for the sparsely sampled patches within the designated area. The utilization of distinct dilation coefficients across the heads fosters the amalgamation of feature information across varying scales, enriching the model's interpretative depth. The computation formula for SWDA is as follows.

$$X = SWDA(Q, K, V, r) \quad (1)$$

In the aforementioned formula, the matrices labeled Q , K , and V correspond to the query, key, and value components, respectively. Each row within these matrices is representative of a distinct feature vector, serving as a foundational element in the computation process.

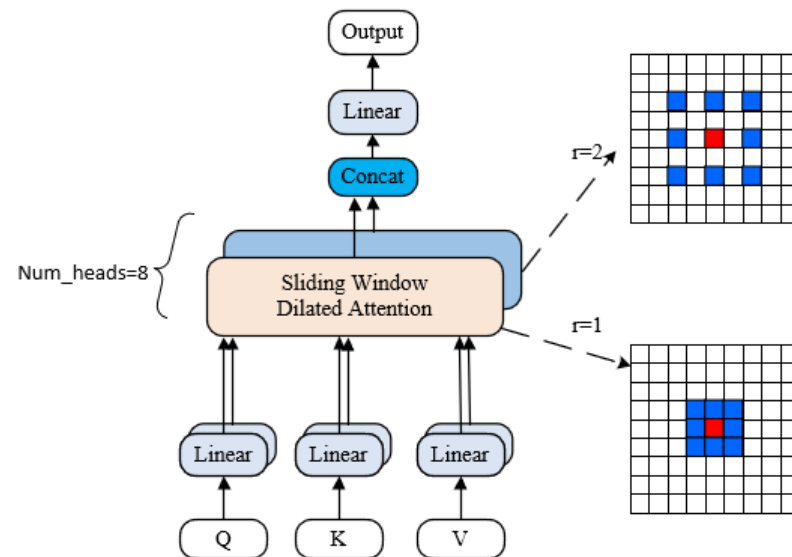


Figure 3. Multi-Scale Dilated Attention (MSDA).

SWDA employs the original feature map's coordinates (i, j) as the focal query point, from which it sparsely selects associated keys and values. It then computes self-attention within a defined sliding window, which is centered on the query point and spans a dimension of $w \times w$. To address the edge positions within the feature map during self-attention computations, SWDA adopts a zero-padding technique, ensuring continuity and preventing data loss. Through this selective and strategic sparse sampling of keys and values around the query point, SWDA not only adheres to the requirements of sparsity but also preserves the principle of locality. This approach is pivotal for effectively capturing dependencies within long-distance sequences, thereby enhancing the model's capability to discern and understand the subtleties of complex spatial-temporal patterns.

3.3. Gradient Flow Information Aggregation Block Col-Depth-Point Convolution (CDPWC)

To minimize the network's overall parameter count while enhancing the precision of detecting passengers' abnormal behaviors, we developed a Gradient Flow Information Aggregation Block, denoted as CDPWC, which utilizes depthwise separable convolutions. Illustrated in Figure 4, the input feature map initially passes through a 1×1 convolutional operation. This key step involves using $c_{out} 1 \times 1$ convolution kernels to perform convolution operations on each pixel position (i, j) of the input feature map $(h \times w \times c_{in})$, outputting c_{out} values. We set $c_{out} = 2 \times c_{in}$, thus doubling the number of channels in the input feature map to optimize the channel size. Through this approach, the shape of the feature map was reshaped, providing a larger space for feature expression and laying the foundation for more complex processing. Subsequent to this channel expansion, the resulting output is bifurcated through a split operation. One segment undergoes processing via a series of Bottleneck_DPWC modules, which are tasked with extracting more sophisticated feature representations. These refined expressions serve to encapsulate the salient feature information of the input data more effectively. The other segment retained from the split is then recombined with the processed features through a concatenation operation. This fusion not only integrates the high-level semantic features with the granular details of the lower-level features but also allows the model to access a spectrum of feature levels

and gradients. Consequently, this rich blend of features significantly contributes to the model's ability to discern and classify abnormal behaviors with greater accuracy. Note that, in Figure 4, h , w , and $c_{in}(c_{out})$ refer to the pixel length, width, and channel dimension of the feature map, respectively; $\times 0.5(n+2)c_{out}$ means that the number of feature maps in the channel dimension is half the number of output channels $c_{out}(n+2)$, where n represents the number of Bottleneck_DPWC blocks.

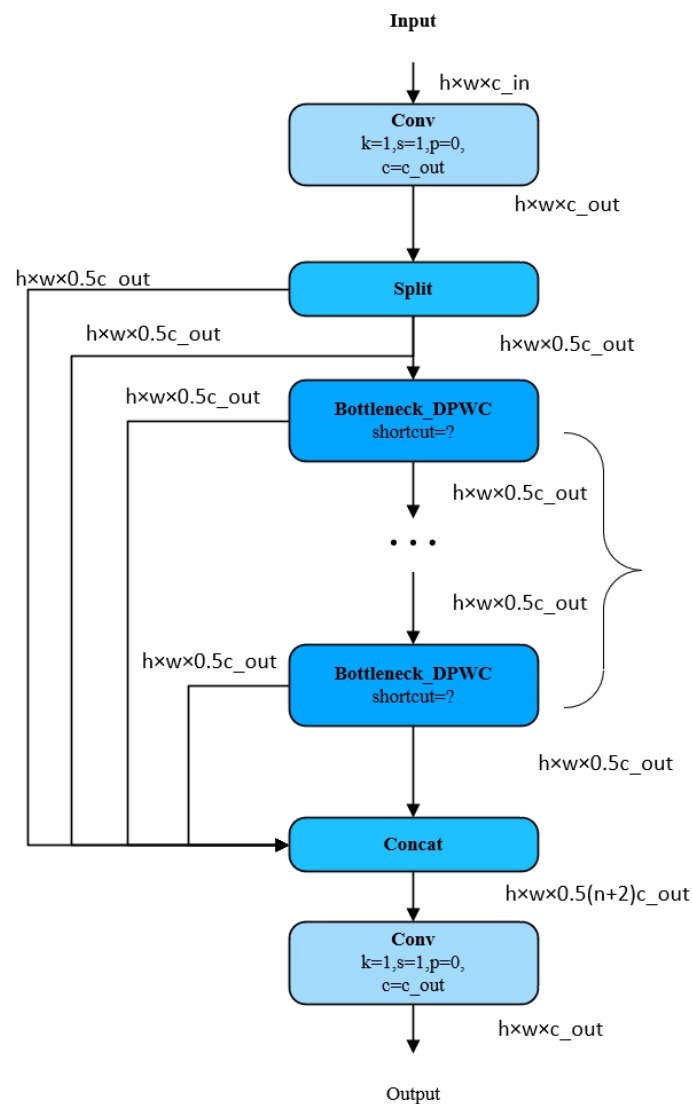


Figure 4. Col-Depth-Point Wise Convolution (CDPWC) module.

Departing from the conventional C2f framework, the Bottleneck_DPWC module we devised incorporates Depthwise Separable Convolution (DSC) [36] to notably diminish the model's parameter while elevating its predictive accuracy. Illustrated in Figure 5, the DSC process is initiated with the input feature map undergoing a DepthWise Convolution(DWC) operation. This entails executing a convolution operation individually on each channel of the feature map, with the outcomes subsequently merged to form the output feature map. Given that each channel is convolved with a singular convolution kernel, this procedure neither expands nor compresses the channel dimension of the feature maps, nor does it effectively capitalize on the inter-channel correlation of features situated at identical spatial coordinates. To foster interaction among the features across different channels, the resultant feature map from DWC is directed through two PointWise Convolution(PWC) blocks. These blocks employ a 1×1 convolution operation to amalgamate the features across channels, thereby augmenting the feature representation's capacity and simultane-

ously reducing the model's parameter count. This strategic enhancement of channel-wise feature interaction substantially bolsters the model's expressive power, making it more adept at accurately modeling complex patterns and relations in the data. The calculation process of the CDPWC module is shown in the following formula, where *BD* represents Bottleneck_DPWC, and + represents *concat* operation.

$$F_{out} = \text{Conv}(\text{Split}(\text{Conv}(\text{input})) + \text{BD}(\text{Split}(\text{Conv}(\text{input})))) \quad (2)$$

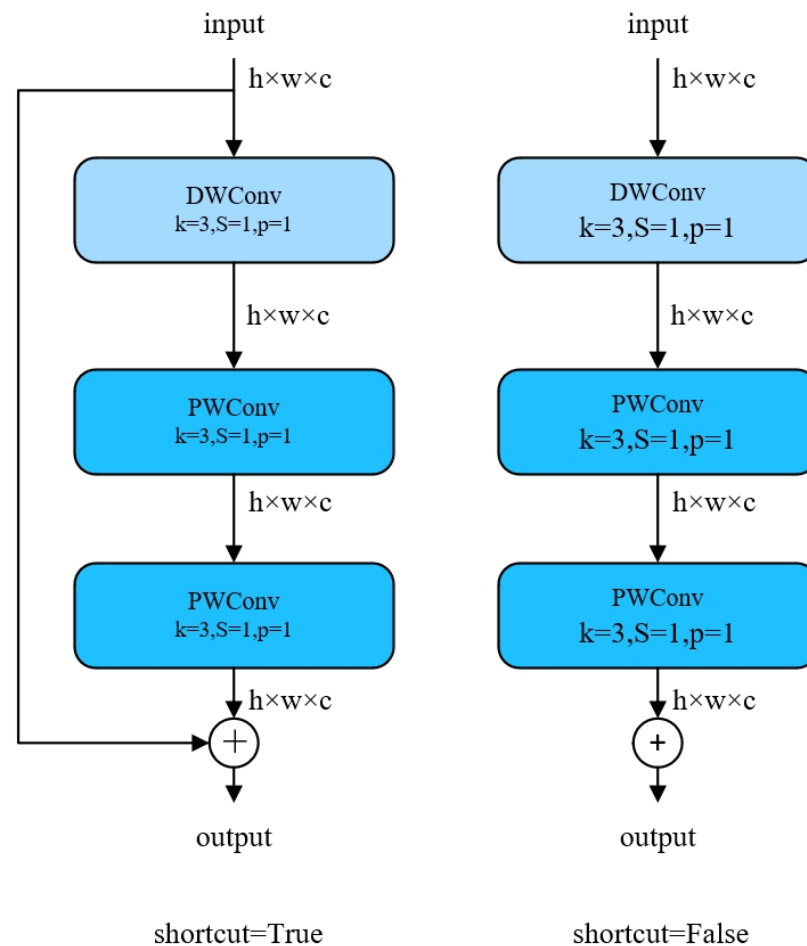


Figure 5. Bottleneck_DPWC structure.

Assume that the channel of the input feature map is M and the size is $D_F \times D_F$; the convolution kernel is $D_K \times D_K$, and the channel of the output feature map is N , while the size is $D_F \times D_F$. Then, the parameters of ordinary convolution are as follows:

$$T1 = D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \quad (3)$$

The parameter amount of depthwise separable convolution is as follows:

$$T2 = D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \quad (4)$$

$$\frac{T2}{T1} = \frac{1}{N} + \frac{1}{D_K^2} \quad (5)$$

From the discussion above, it becomes apparent that Depthwise Separable Convolution (DSC) significantly outperforms traditional convolution in terms of computational efficiency.

4. Experiment

4.1. Experimental Environment and Parameter Settings

The experimental setup was conducted on a server boasting the Ubuntu-20.04 operating system, equipped with an NVIDIA GeForce RTX 3090 GPU for graphical processing. Model training was executed utilizing the PaddlePaddle deep learning framework. Within the training hyperparameters, Momentum was selected as the optimization algorithm, with a set momentum coefficient of 0.9. The learning rate was scheduled to decrease following a cosine annealing strategy. All models were trained from scratch, without the aid of pre-trained weights, and adhered uniformly to a consistent data augmentation protocol. Additional training parameters were retained at the framework's default settings for consistency across the experiments. Table 1 lists in detail the training parameters adopted in our proposed model for further reference. And our code is available at <https://github.com/Bradly-s/TSM-CDPMSDANet> (accessed on 17 June 2024).

Table 1. Experimental parameter settings.

Experimental Parameter	Setting
Batch size	2
Epochs	80
warmup_epochs	10
warmup_start_lr	0.002
cosine_base_lr	0.001
Optimizer	Momentum
Momentum	0.9
weight_decay	L2

4.2. Elevator Passenger Abnormal Behavior Dataset

For the examination of abnormal passenger behaviors in indoor elevator scenarios, we amassed video footage covering four distinct categories: door picking, jumping, kicking, and door blocking. From each video capturing these abnormal behaviors, segments ranging from 4 to 10 s were extracted. Subsequently, individual frames were isolated from these segments to compile corresponding images. These images were then organized into a dataset formatted for experimental use. The cumulative count of images assembled for the respective categories of abnormal behaviors are as follows: door picking yielded 4334 images, jumping activities contributed 2891 images, instances of kicking resulted in 4775 images, and door blocking accounted for 6311 images. Data examples are shown in Figure 6.

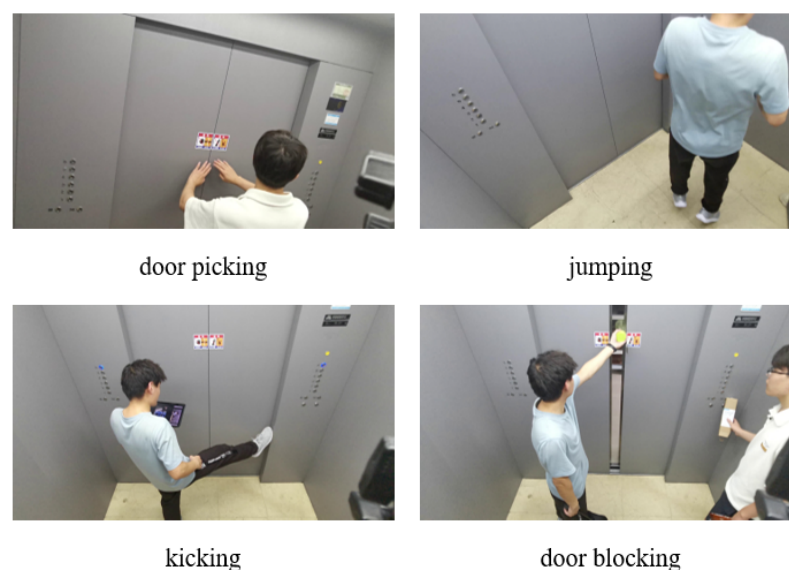


Figure 6. Examples of four types of abnormal behavior.

4.3. Public Behavior Recognition Dataset

In order to assess the generalization capabilities and robustness of our model, we carried out evaluative tests on four publicly available datasets. The datasets utilized for this purpose are delineated as follows:

(1) UCF101 dataset

UCF101 [37] is a comprehensive action recognition dataset that encompasses 101 distinct categories of actions, featuring roughly 100 videos per category. It stands out due to its unparalleled diversity in action types, coupled with significant variations in several aspects, including camera movement, the appearance and pose of objects, the scale of objects, viewpoints, the presence of cluttered backgrounds, and lighting conditions. These factors collectively render UCF101 the most challenging dataset in the field to date.

(2) UCF24 dataset

UCF10_24 represents a specialized subset of the broader UCF101 dataset, distinguished by its utilization of an alternative labeling scheme. In this subset, each video is characterized by the presence of, at most, a single type of target behavior, and bounding boxes (bboxes) are employed exclusively to annotate individuals engaged in the specified target behavior.

(3) hmdb51 dataset

The HMDB-51 [38] dataset comprises 51 distinct categories of behaviors, with each category containing a minimum of 101 video instances, summing up to a total of 6766 short video clips. The behaviors captured in these videos are further classified into five categories: actions involving facial expressions and object interactions, fundamental body movements, actions that involve interacting with objects, and complex bodily actions. In our experimental framework, for each category within the dataset, 70% of the videos were designated for the training set, while the remaining 30% were allocated to the test set.

(4) Something-Something-v1 dataset

The Something-Something-v1 [39] video dataset is made available through a substantial TGZ archive, systematically segmented into portions of up to 1 GB, culminating in a total download volume of 25.2 GB. It encompasses a comprehensive collection of 108,499 videos, each comprising JPG images that maintain a consistent height of 100 pixels and a variable width. These JPG images are meticulously extracted from the original footage at a rate of 12 frames per second. The dataset is organized into a structured format, featuring a training set with 86,017 images, a validation set comprised of 11,522 images, and a test set that includes 10,960 images.

5. Results and Discussions

5.1. Evaluation Indicators

Our video classification model underwent a thorough evaluation process using a dataset specifically composed of four categories of abnormal elevator passenger behaviors, in addition to four publicly available datasets. To determine the model's precision, we employed the classification accuracy metric (Top-1 accuracy, abbreviated as Acc). The formula employed to calculate the accuracy is articulated below:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \quad (6)$$

5.2. Comparative Experiment

Within the dataset detailing abnormal passenger behaviors in elevators that we curated, our model underwent comparative analysis against leading-edge models in the field. The outcomes of this comparison are detailed in Table 2. An examination of the table reveals that our model surpasses contemporary advanced models in terms of the Top-1 accuracy (Acc) metric. Specifically, when juxtaposed with the PPTSM model, our model

demonstrated a significant enhancement of 10%, achieving a Top-1 accuracy of 95%. Even when compared to the more accurate SlowFast model, our model exhibited an improvement of 3%, underscoring the exemplary performance of our approach.

Table 2. Comparative experiments on abnormal behavior dataset of passengers in elevators.

Methods	Top1 Acc	Parameter Size (M)	GFLOPs	Model Size (MB)
PPTSM	0.85	23.585	34.809	90.02
MoViNet	0.63	1.907	61.029	21.73
TSN	0.8	23.571	32.876	89.73
PP-TSN	0.9	23.591	34.809	89.81
TSM	0.65	23.571	32.876	89.73
Ours	0.95	32.857	33.852	125.40

To ascertain the generalizability and robustness of our model, we conducted a series of experiments across four public behavior recognition datasets: UCF24, UCF101, HMDB51, and Something-Something-v1. We evaluated both our enhanced model and the baseline model in these experiments. A comparative analysis, focusing on the Top-1 accuracy (Acc) metric, is systematically presented in Table 3.

Table 3. Comparative experiments on public behavior recognition dataset.

Dataset	PPTSM	Ours
UCF24	0.6030	0.6710
UCF101 (split1)	0.4970	0.5348
UCF101 (split2)	0.4819	0.5302
UCF101 (split3)	0.4721	0.5695
HMDB51	0.2778	0.4902
Something-Something-v1	0.2380	0.2776

Table 3 clearly demonstrates the performance enhancements achieved using our improved model over the original model. Specifically, with the UCF24 dataset, there was a significant increase in Top-1 accuracy (Acc) by 6.8%. Our model improved by 3.78%, 4.83%, and 9.74%, respectively, on the three divisions, split1, split2, and split3, of the UCF101 data set. With the HMDB51 dataset, our model showcased a remarkable improvement of 21.24% compared to the original model. Conversely, with the Something-Something-v1 dataset, the enhancement was more modest at 3.959%. The relatively smaller gain with the Something-Something-v1 dataset can be attributed to its focus on the interaction between objects, for which understanding actions often relies heavily on contextual information. Given that such context is somewhat limited when extracted solely from video clips, this could explain the lesser degree of improvement. The results from these experiments across various public datasets affirm the robustness and generalization capability of our model, indicating its strong performance even in scenarios requiring nuanced contextual interpretation.

5.3. Ablation Experiment

To assess the effectiveness of our proposed model, we undertook a series of ablation studies with four distinct datasets of abnormal passenger behavior, namely door picking, jumping, kicking, and door blocking. These experiments were designed to evaluate the efficacy of the Multi-Scale Dilated Attention (MSDA) we integrated, as well as the performance of the proposed Gradient Flow Information Aggregation Block (CDPWC). Additionally, these studies aimed to confirm whether the Gradient Flow Information Aggregation Block could indeed effectively minimize the model's parameter count. Employing the video classification model PPTSM as our baseline, we incrementally introduced our modules to conduct these ablation studies. The outcomes of these experiments are meticulously detailed in Table 4.

The data presented in the table clearly indicate that, upon integrating the C2f module, there was a notable 5% increase in accuracy, accompanied by a reduction in the model's parameters by 2.515 M and a decrease in model size by 9.61 MB. However, this was offset by a rise in computational demand, evidenced by an increased need for 3.006 GFLOPs. When the proposed CDPWC module was added to the baseline independently, it also elevated model accuracy by 5%. More impressively, it resulted in a substantial decrease in model parameters of 7.507 M, a significant reduction in computational load of 7.533 GFLOPs, and a notable decline in model size of 28.63 MB. The reason for the decrease in the parameter count is that, in the CDPWC module, we replaced the original two ordinary 3×3 convolutions with depthwise separable convolutions in Bottleneck_DPWC. According to Equation (5), the computational complexity of the model significantly decreased. The standalone inclusion of the MSDA module led to a significant boost in accuracy of 10%, but this came at the cost of an increase in model parameters of 16.785 M, a surge in computational requirements of 6.576 GFLOPs, and an enlargement of the model size from 90.02 MB to 154.03 MB. From the comparisons drawn in Table 4, it can be inferred that, while the C2f module enhances model accuracy and reduces both the number of parameters and the model size, it does so at the expense of increased computational complexity. On the other hand, adding the CDPWC module not only effectively improves the model's accuracy but also markedly diminishes both the parameter count and the model size, thereby reducing the overall model complexity. The MSDA leverages varying dilation rates to expand the receptive field of the convolutional operations, affording it the capacity to discern intricate feature details associated with different behaviors. Consequently, the inclusion of MSDA enhances model accuracy and, due to its transformer-like structure, amplifies model complexity. In this study, we synergized both the CDPWC and MSDA modules. With this combination, at an accuracy peak of 95%, we were able to effectively capture subtle characteristics of distinct behaviors while simultaneously bolstering the model's ability to identify a variety of abnormal behaviors.

Table 4. Comparison of evaluation of each module in the ablation experiment.

Methods	Top1 Acc	Parameter Size (M)	GFLOPs	Model Size (MB)
baseline	0.850	23.585 M	34.809	90.02
baseline+C2f	0.900	21.070 M	37.815	80.41
baseline+CDPWC	0.902	16.078 M	27.276	61.39
baseline+MSDA	0.950	40.370 M	41.385	154.03
baseline+CDPWC+MSDA	0.951	32.857 M	33.852	125.34

5.4. Visualization

To vividly illustrate the comparative performance between our model and the PPTSM model, we plotted the prediction accuracies of both models across different categories within the UCF24 and HMDB51 datasets. These comparisons are visually represented in Figures 7 and 8. From Figure 7, it is evident that, within the UCF24 dataset, our model demonstrates a marginal decrease in accuracy for the categories of "GolfSwing," "Skiing," and "SoccerJuggling." Despite these specific instances, our model outperforms the PPTSM model in the majority of categories, showcasing superior overall results. Figure 8 further supports our model's enhanced performance across the board within the HMDB51 dataset. Despite minor declines in the "smile" and "walk" categories, these do not detract from the overall superiority of our improved model. This evidence underscores the enhanced efficacy and applicability of our model across a diverse range of categories.

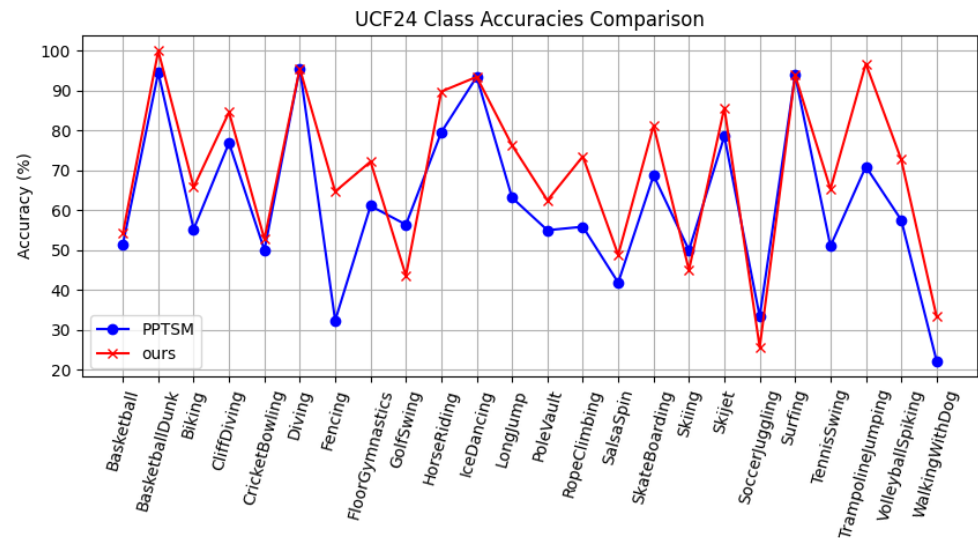


Figure 7. Category accuracy compared with PPTSM in UCF24 dataset.

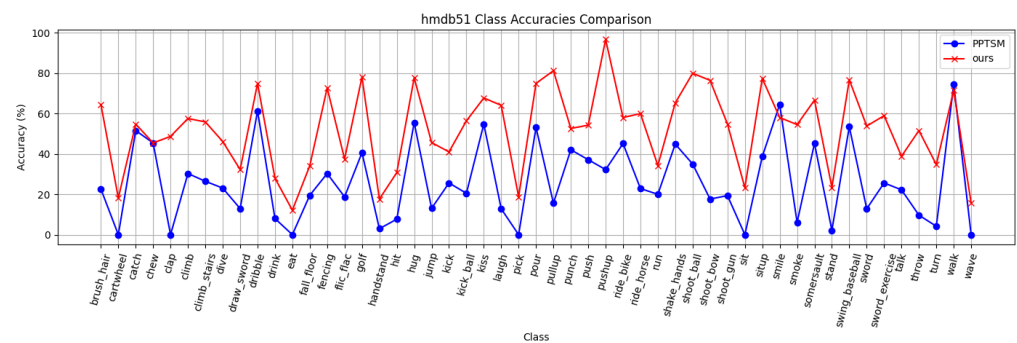


Figure 8. Category accuracy compared with PPTSM in HMDB51 dataset.

6. Conclusions

In this paper, we have embarked on an in-depth exploration of the intricate domain of the detection of passengers' abnormal behavior in elevators. Therefore, we collected data on four abnormal behaviors of passengers in elevator cabins (door picking, jumping, kicking, and door blocking) from real-life scenarios and established a proprietary dataset as the data support for this study. Our goal was to provide a robust solution that aptly addresses the limitations imposed via model accuracy inherent to the scenarios of human behavior detection in elevators. Accordingly, we have proposed the TSM-CDPMSDANet model. In our proposed model, the multi-scale dilated attention (MSDA) mechanism is introduced to capture the subtle properties of various abnormal behaviors and enable the model to effectively identify passengers' behaviors at different scales. Another important module is the gradient flow information aggregation block Col-Depth-Point Convolution (CDPWC), which is designed to promote the fusion of complex, low-level detail features and high-level semantic features and ultimately achieve a synergistic effect of feature complementarity.

The experimental results demonstrate that our model achieved 95% accuracy in identifying abnormal behavior with our established dataset. Additionally, when compared to various advanced models and evaluated across different public datasets, our model consistently outperforms, demonstrating robust generalization capabilities. Looking to the future, our work lays the groundwork for practical applications in human behavior detection, offering enhanced accuracy. However, due to practical constraints (the fixed-angle camera systems in the elevators around us and the insufficient number of passengers to create occlusive conditions), we cannot capture images of passengers' occlusions or multi-angle views inside the elevator. Consequently, we aim to address this issue in the future.

Author Contributions: Conceptualization, W.S. and S.Y.; methodology, J.L., W.S., Y.F., N.Y. and S.Y.; validation, J.W., W.S., J.L. and N.Y.; formal analysis, Y.F., N.Y. and S.Y.; investigation, Y.F. and S.Y.; resources, J.L.; data curation, J.L.; writing—original draft preparation, J.L., W.S. and Y.F.; writing—review and editing, J.L., J.W. and S.Y.; visualization, W.S. and Y.F.; supervision, J.L. and S.Y.; project administration, J.L.; funding acquisition, J.L., J.W. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Zhejiang Provincial Key Research and Development Project (No. 2024C01135) and Scientific Research Fund of Zhejiang Provincial Education Department (No. Y202352150).

Data Availability Statement: The study's data are available on <https://github.com/Bradly-s/TSM-CDPMSDANet> (accessed on 17 June 2024).

Acknowledgments: We would like to acknowledge Huijie Zhu's valuable contribution to this article.

Conflicts of Interest: Author Jingsheng Lei was employed by the company Zhejiang Xinzailing Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wang, H.; Zeng, M.; Xiong, Z.; Yang, F. Finding main causes of elevator accidents via multi-dimensional association rule in edge computing environment. *China Commun.* **2017**, *14*, 39–47. [\[CrossRef\]](#)
2. Lan, S.; Gao, Y.; Jiang, S. Computer vision for system protection of elevators. *J. Phys. Conf. Ser.* **2021**, *1848*, 012156. [\[CrossRef\]](#)
3. Prahlow, J.A.; Ashraf, Z.; Plaza, N.; Rogers, C.; Ferreira, P.; Fowler, D.R.; Blessing, M.M.; Wolf, D.A.; Graham, M.A.; Sandberg, K.; et al. Elevator-related deaths. *J. Forensic Sci.* **2020**, *65*, 823–832. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Prabha, B.; Shanker, N.; Priya, M.; Ganesh, E. A study on human abnormal activity detecting in intelligent video surveillance. In Proceedings of the International Conference on Signal Processing & Communication Engineering, Andhra Pradesh, India, 11–12 June 2021; AIP Publishing: Melville, NY, USA, 2024; Volume 2512. [\[CrossRef\]](#)
5. Li, N.; Ma, L. *Typical Elevator Accident Case: 2002–2016*; China Labor and Social Security Publishing House: Beijing, China, 2019; p. 001.
6. Zhu, Y.; Wang, Z. Real-time abnormal behavior detection in elevator. In Proceedings of the Intelligent Visual Surveillance: 4th Chinese Conference, IVS 2016, Beijing, China, 19 October 2016; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2016; pp. 154–161. [\[CrossRef\]](#)
7. Sun, Z.; Xu, B.; Wu, D.; Lu, M.; Cong, J. A real-time video surveillance and state detection approach for elevator cabs. In Proceedings of the 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), IEEE, Chengdu, China, 23–26 October 2019; pp. 1–6. [\[CrossRef\]](#)
8. Liu, S.; An, Z.; Wang, N.; Bai, D.; Yu, X. Research on elevator passenger fall detection based on machine vision. In Proceedings of the 2021 3rd International Conference on Advances in Civil Engineering, Energy Resources and Environment Engineering, Qingdao, China, 28–30 May 2021; Volume 791, p. 012108. [\[CrossRef\]](#)
9. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [\[CrossRef\]](#)
10. Lan, S.; Jiang, S.; Li, G. An elevator passenger behavior recognition method based on two-stream convolution neural network. In Proceedings of the 2021 4th International Symposium on Big Data and Applied Statistics (ISBDAS 2021), Dali, China, 21–23 May 2021; Volume 1955, p. 012089. [\[CrossRef\]](#)
11. Chen, Y.; Zhao, Q.; Fan, Q.; Huang, X.; Wu, F.; Qi, J. Falling Behavior Detection System for Elevator Passengers Based on Deep Learning and Edge Computing. In Proceedings of the 2nd International Conference on Electronics Technology and Artificial Intelligence (ETAI 2023), Changsha, China, 18–20 August 2023; Volume 2644, p. 012012. [\[CrossRef\]](#)
12. Shi, Y.; Guo, B.; Xu, Y.; Xu, Z.; Huang, J.; Lu, J.; Yao, D. Recognition of abnormal human behavior in elevators based on CNN. In Proceedings of the 2021 26th International Conference on Automation and Computing (ICAC), IEEE, Portsmouth, UK, 2–4 September 2021; pp. 1–6. [\[CrossRef\]](#)
13. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7083–7093.
14. Jiao, J.; Tang, Y.M.; Lin, K.Y.; Gao, Y.; Ma, A.J.; Wang, Y.; Zheng, W.S. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Trans. Multimed.* **2023**, *25*, 8906–8919. [\[CrossRef\]](#)
15. Gall, J.; Lempitsky, V. Class-specific hough forests for object detection. In *Decision Forests for Computer Vision and Medical Image Analysis*; Springer: London, UK, 2013; pp. 143–157. [\[CrossRef\]](#)
16. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.

17. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576. [\[CrossRef\]](#)
18. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
19. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Process. Image Commun.* **2019**, *71*, 76–87. [\[CrossRef\]](#)
20. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
22. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541. [\[CrossRef\]](#)
23. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200. <https://doi.org/10.48550/arXiv.1711.08200>.
24. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
25. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference On Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [\[CrossRef\]](#)
26. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12026–12035.
27. Yin, M.; He, S.; Soomro, T.A.; Yuan, H. Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network. *Expert Syst. Appl.* **2023**, *226*, 120080. [\[CrossRef\]](#)
28. Feng, S.; Niu, K.; Liang, Y.; Ju, Y. Research on elevator intelligent monitoring and grading warning system. In Proceedings of the 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 24–26 September 2021; pp. 145–148. [\[CrossRef\]](#)
29. Zhao, J.; Yan, G. Passenger Flow Monitoring of Elevator Video Based on Computer Vision. In Proceedings of the 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 3–5 June 2019; pp. 2089–2094. [\[CrossRef\]](#)
30. Wu, D.; Wu, S.; Zhao, Q.; Zhang, S.; Qi, J.; Hu, J.; Lin, B. Computer vision-based intelligent elevator information system for efficient demand-based operation and optimization. *J. Build. Eng.* **2024**, *81*, 108126. [\[CrossRef\]](#)
31. Qi, Y.; Lou, P.; Yan, J.; Hu, J. Surveillance of abnormal behavior in elevators based on edge computing. In Proceedings of the 2019 International Conference on Image and Video Processing, and Artificial Intelligence, Shanghai, China, 23–25 August 2019; Volume 11321, p. 1132114. [\[CrossRef\]](#)
32. Shu, G.; Fu, G.; Li, P.; Geng, H. Violent behavior detection based on SVM in the elevator. *Int. J. Secur. Appl.* **2014**, *8*, 31–40. [\[CrossRef\]](#)
33. Jia, C.; Yi, W.; Wu, Y.; Huang, H.; Zhang, L.; Wu, L. Abnormal activity capture from passenger flow of elevator based on unsupervised learning and fine-grained multi-label recognition. *arXiv* **2020**, arXiv:2006.15873. <https://doi.org/10.48550/arXiv.2006.15873>.
34. Wang, Z.; Shen, Z.; Chen, J.; Li, J.; Wu, W. Recognition of Abnormal Behaviors of Elevator Passengers Based on Temporal Shift and Time Reinforcement Module. In Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC), Dalian, China, 27–29 July 2023; pp. 670–675. [\[CrossRef\]](#)
35. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 558–567.
36. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. <https://doi.org/10.48550/arXiv.1704.04861>.
37. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402. <https://doi.org/10.48550/arXiv.1212.0402>.
38. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, IEEE, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563. [\[CrossRef\]](#)
39. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The “something something” video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5842–5850. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.