Study on Human Hazardous Behavior Recognition and Monitoring System in Slide Facilities Based on Improved HRNet Network

Chen Chen, Huiyu Xiang*, Song Huang*, Yanpei Zhang
School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing, China

***Abstract*—In recent years, accidents involving slide playground equipment have frequently occurred due to various reasons, attracting significant attention. Reducing or even eliminating these accidental injuries has become an urgent technical issue to address. Currently, the safety management of slide playground facilities still relies on manual monitoring, and the level of technology for detecting and intelligently recognizing hazardous behaviors on slides needs improvement. This paper proposes a behavior detection system based on human skeleton sequence information to address the issue of recognizing hazardous behaviors on slides. To resolve the feature fusion loss problem that arises when HRNet extracts feature information from images of different resolutions, this paper introduces a Flow Alignment Module (FAM) and an Attention-aware Feature Fusion (AFF) module to improve the network structure. Experimental results show that the improved skeleton sequence extraction model exhibits good computational efficiency and accuracy on the dataset, achieving an accuracy rate of over 90%. The human behavior recognition system proposed in this paper effectively meets detection requirements, providing new technical assurance for the safe use of slide playground equipment.**

***Keywords*—Playground equipment; object detection; skeleton sequence; flow alignment module; human behavior recognition**

I. INTRODUCTION

Slide playground equipment is a common feature in parks, shopping malls, and large communities, beloved by children, and playing a crucial role in their growth and development [1]. Evaluating the safety of large sliding playground equipment typically involves analyzing various aspects such as equipment, personnel, management, and the environment [2]. Heinrich [3] discovered through investigation that the majority of known safety accidents are caused by human hazardous behaviors. The impact of external environments on these facilities and their safety issues has always been a significant concern.

Wenxiang Cui [4] analyzed 913 cases of accidental injuries among children in kindergartens, examining the causes of safety accidents, the level of awareness regarding accidental injuries, and the behavioral characteristics prone to accidents. The results showed that each child has unique personality traits, which influence their behavior. Children who choose high-risk behaviors are more likely to cause safety accidents. Although research on the safety of slide playground equipment has made some progress, there are still deficiencies. Currently, the detection of dangerous behaviors during the operation of slides mainly relies on manual observation. Due to the immature mental development of children, they lack awareness of dangerous behaviors. Additionally, parents find it difficult to monitor their children throughout the entire play process, making it easy for dangerous behaviors to go unnoticed and unaddressed, leading to accidents.

To enhance the safety of slide playground equipment, deep learning-based intelligent detection technology can analyze and process large amounts of data, training recognition models to identify behaviors that may lead to safety accidents, effectively preventing such incidents. In practical applications, it is crucial to continuously optimize algorithms and monitoring systems to improve accuracy and predictive effectiveness, enhancing the informatization level of safety management for playground equipment. This helps children develop correct safety habits, thereby reducing the occurrence of safety accidents.

Currently, methods for recognizing dangerous behaviors mainly include manual inspection, wearable sensors, and computer vision techniques [5]. Studies have shown that using human pose information can aid in target recognition. For example, Guo [6] proposed a method that uses human skeletal information for real-time recognition, simplifying dynamic movements into static poses and matching these poses with a database of dangerous behaviors, thereby reducing misjudgments in complex environments. Yang Bin [7] combined target detection with skeleton point extraction technology. They used human skeletal information to determine initial behavior categories and target recognition technology to locate phones and cigarettes, assessing whether dangerous behaviors occurred based on the relationship between the person and the detected target. Wang Hong [8] used the OpenPose algorithm to extract skeleton diagrams of personnel in electric power operation sites and utilized the VGG network to extract feature information from all obtained skeleton diagrams, providing a framework for combining pose estimation and deep learning techniques. Zhang [9] proposed a two-stage skeleton-RGB integrated model for predicting human actions in human-robot collaborative assembly, improving prediction accuracy and efficiency for highly similar human actions.

Some studies have used target recognition results as inputs for behavior recognition. Han and Lee [10] combined human skeletal information with 3D reconstruction algorithms, converting 2D skeletal information into actual-sized 3D models, achieving precise descriptions and restorations of workers' actions. Xiong Ruoxin [11] analyzed the actions of construction workers using 3D pose estimation. However, the datasets used

in these studies are difficult to obtain and mostly collected in laboratories, leading to insufficient generalization of the training models to complex real-world environments and poor recognition performance. Fu [12] proposed obtaining preliminary image frame information through target detection and then inputting this into a lightweight OpenPose network to obtain real-time coordinates of human skeletal key points. Combining the two techniques can enhance the speed of skeletal key point extraction networks, and by calculating the set central point coordinates of selected skeletal key points, determining whether a person has fallen based on the descent speed and the human aspect ratio. Takkar [13] proposed a part-based graph convolutional network that first performs graph convolution in subgraphs constructed for each body part, then propagates information between subgraphs through shared nodes. However, this method has limited capability for part-level information modeling. Huang [14] used relationship modules and attention modules to learn the correlations and importance between body parts and used unpooling operations to bridge part-level and joint-level graphs to capture rich motion information. However, for some fine-grained actions, the recognition performance may be limited because the cooperation of body parts is not obvious, and unpooling operations may weaken or even obscure joint-level information. Qiu [15] proposed a new multi-granularity fragment focus network (MGCF-Net), achieving good performance on two large-scale benchmarks for skeleton-based action recognition. Wu [16] mapped skeletal data to multiple granularities, using graph convolution and self-attention mechanisms to capture relevant information at each granularity and using weighted summation to integrate multi-granularity information. Jianbao Zhu [17] used the Canny operator to process images collected at construction sites and employed the Hough line detection algorithm to detect lines in edge binary images and calibrate them. They used a human skeletal key point extraction algorithm to obtain the coordinates of the feet in the images, determining whether workers were in safe areas based on the positions of their feet.

However, the current feature fusion methods usually implement linear operations such as summation and concatenation, which cannot effectively integrate features of different resolutions, resulting in semantic information loss and errors. Additionally, this increases the computational load during network feature fusion, reducing the speed and accuracy of human skeletal key point extraction algorithms. Increasingly, studies are adopting deep learning models to handle more complex behavior recognition tasks. These methods use end-to-end training to enable the models to directly learn features from raw data, avoiding the cumbersome process of manually designing features and preprocessing, thus improving recognition efficiency. Moreover, these models better understand the correlations between behaviors, enhancing overall performance. As hardware performance continues to improve and algorithms are continuously optimized, more studies are focusing on improving the real-time performance and efficiency of behavior recognition methods that combine target recognition and pose estimation, further strengthening the technical support for intelligent environmental perception and behavior understanding. The research route of this paper is shown in Fig. 1.

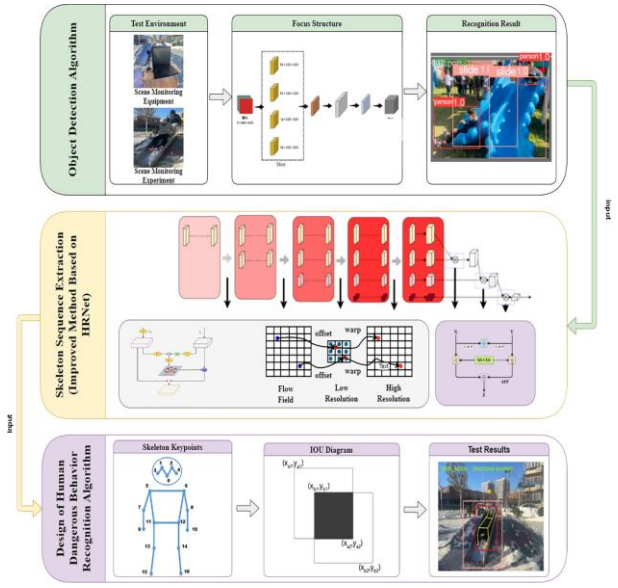


Fig. 1. Technical route diagram.

The main contributions of this study can be summarized as follows:

Development of a Human Pose Estimation Algorithm: A novel algorithm capable of integrating semantic information from images of varying resolutions was constructed. Through careful design and optimization, this algorithm enables real-time and high-performance human behavior recognition on a computer platform.

Proposition of an Improved High-Resolution Network (HRNet) Scheme: An improved HRNet scheme was proposed to address the pixel information loss during up-sampling and down-sampling of multi-resolution images, and the insufficient semantic information fusion during feature exchange of different resolution images. This enhancement significantly improved the accuracy of human skeleton key point extraction and the stability of logical judgment in the code.

Utilization of Target Recognition Information: Information from human bounding boxes and the number and spatial relationships of skeleton key points were utilized to describe action behavior categories such as crowding, close proximity, and staying still. This approach facilitated the establishment of multiple hazardous behavior recognition models.

Experimental Validation: Experimental results indicated that the developed human behavior recognition model exhibited outstanding performance in identifying hazardous incidents on slides. The testing outcomes demonstrated that this human behavior detection system meets the requirements for hazardous behavior detection, providing new technical assurance for the design and operational safety of playground equipment.

The structure of this paper is organized as follows: First, the Introduction in Section I presents the background and research significance of hazardous behavior recognition in slide facilities, along with a review of related research. Next, the Materials and Methods in Section II provides a detailed description of the experimental data sources, the selection and performance evaluation of object detection algorithms, and the skeleton

sequence extraction method based on the improved HRNet network. Subsequently, the Results and Discussion in Section III presents the experimental results of the improved algorithm and provides an in-depth analysis of its performance. Finally, the Conclusion in Section IV summarizes the main contributions of this study and proposes directions for future research. Through this structure, the paper aims to provide an intelligent detection solution based on deep learning for the safety management of slide facilities, effectively reducing accidental injuries among children during slide usage.

II. Materials and Methods

A. Experimental Data

The image dataset used in this study was sourced from web scraping and the video surveillance systems of amusement facilities. This dataset contains 5,000 images, covering various angles of slides, individuals on slide facilities, and people around the slides. For model training and evaluation, we divided the dataset into training and testing sets, with 4,500 images in the training set and 500 images in the testing set. We used the labelimg package to annotate the dataset, categorizing it into two classes: person and slide, and generating .txt format files required for YOLOv5 training. Fig. 2 shows some example images.

Fig. 2. Image data (Parental Consent Obtained).

The training settings parameters include: initial learning rate (Learning_rate) of 0.001, number of epochs (Epoch) set to 500, batch size (Batch Size) of 4, and momentum factor (Momentum) of 0.9.

B. Selection and Performance Evaluation of Object Detection Algorithms

Single-stage object detection algorithms have significant advantages in terms of computation speed and real-time performance, as they predict object categories and locations through an end-to-end network structure. This makes them suitable for scenarios with high real-time requirements, such as monitoring and security systems. Currently, the YOLO (You Only Look Once) series of object detection algorithms are among the most mature applications, formalizing the object detection problem as a regression problem.

Compared to previous versions, YOLOv5 adds a Focus structure for image slicing, reducing parameter and computation amounts while improving detection accuracy and speed. Its handling of targets of different scales is also stronger [18]

YOLOv5 is built on a neural network model, primarily using a CNN model as its backbone network, combined with bounding box and confidence predictions to achieve accurate object detection. In terms of optimization and improvements, YOLOv5 makes meticulous adjustments compared to its predecessor YOLOv4 in the input end, backbone network, neck network, and loss functions. Specifically, it introduces the Focus and CSP structures to enhance feature extraction and network learning capabilities. The neck employs an FPN+PAN structure to achieve multi-scale feature fusion, further improving detection precision and efficiency. YOLOv5 uses multiple loss functions to optimize classification, localization, and confidence predictions.

YOLOv5 incorporates a Focus structure and adaptive image scaling at the input end. Traditional object detection algorithms typically scale raw images to a unified size for network input processing. However, this scaling can introduce black borders of varying sizes at the image edges, leading to information redundancy. These extra pixels do not contain useful target information, potentially increasing the model's computational load and reducing inference speed. YOLOv5 minimizes these black borders, enhancing network computation speed. Additionally, while YOLOv4 uses the CSP structure only in the backbone network, YOLOv5 employs two CSP structures: CSP1_X in the backbone network and CSP2_X in the neck, enhancing the network's feature fusion capability. CSPNet can reduce network computation without significantly impacting accuracy [19].

The Focus structure performs further feature extraction, with a core step being the slice operation [20]. The initial image, sized 640×640×3, is sliced into a 320×320×12 image, then convolved with 32 kernels to produce a 320×320×32 feature map. The data is divided into four parts, each equivalent to a 2x down-sampled version, concatenated along the channel dimension, and then convolved. CSP structure is shown in Fig. 3.

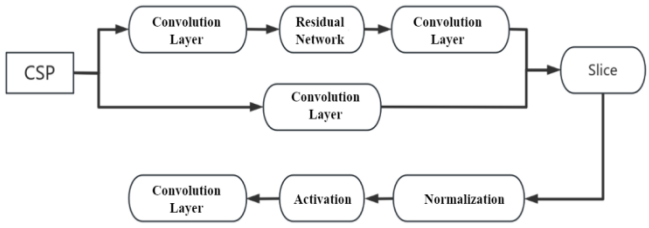
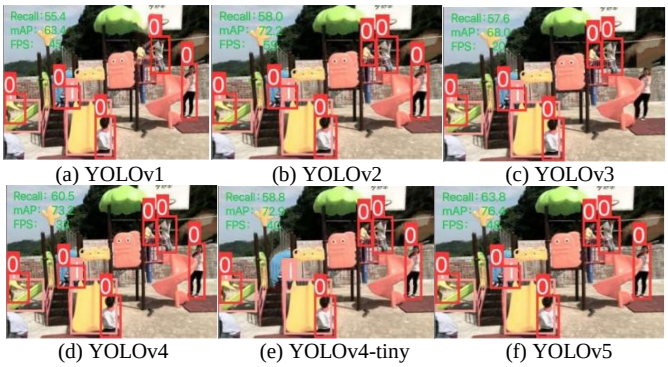


Fig. 3. Cross-stage partial (CSP) structure.

The goal of this paper is to perform real-time behavior recognition in multi-person environments. To ensure real-time effectiveness during the recognition process, a network must be selected that satisfies both fast detection and high detection accuracy requirements. This study conducted tests on various versions of the YOLO series algorithms within single-stage object detection algorithms, using datasets collected through Python web scraping techniques. Performance was measured using three metrics: frames per second (FPS), recall rate, and mean average precision (mAP). The test results for different versions of the YOLO series algorithms are shown in Table I, and the detection results for human bodies are illustrated in Fig. 4.

TABLE I TEST RESULTS OF DIFFERENT YOLO SERIES ALGORITHMS

Algorithm Name	Network Structure	Recall (%)	mAP (%)	FPS
YOLOv1	GoogLeNet	55.4	63.4	45
YOLOv2	Dark Net-19	58.0	72.2	47
YOLOv3	Dark Net-53	57.6	68.0	20
YOLOv4	CSP Dark Net53	60.5	73.2	33
YOLOv4-tiny	CSP Dark Net53-tiny	58.8	72.9	40
YOLOv5	CSP Dark Net53(Focus)	63.8	76.4	48



(a) YOLOv1 (b) YOLOv2 (c) YOLOv3

(d) YOLOv4 (e) YOLOv4-tiny (f) YOLOv5

Fig. 4. Detection results of different YOLO series algorithms.

According to the detection result images, all six YOLO detection networks accurately identify the two categories of people and slides in the images, precisely bounding their positions and sizes. The table data shows that each network performs differently in terms of detection performance. The YOLOv2 and YOLOv5 models have the fastest detection speeds, while the YOLOv4 and YOLOv5 models exhibit the highest target prediction accuracy. On the given dataset, the YOLOv5 model performs the best among multiple detection models. The test results clearly indicate that the YOLOv5 network is most suitable for the real-time detection needs of this study. Therefore, this paper will adjust the parameters of the human and slide detection algorithms based on the YOLOv5 model.

C. Skeleton Sequence Extraction Method Based on Improved Feature Fusion Module

In the field of computer vision, target tracking plays a crucial role by modeling the shape and movement trajectory of targets and using data association methods to achieve continuous tracking of targets in video streams. To achieve this, information from adjacent frames is typically utilized [21]. In multi-target tracking scenarios, there may be multiple targets in the video sequence with similar shapes and movement trajectories, necessitating the use of multi-target tracking algorithms to ensure continuous tracking of each target. In slide playground facilities, it is essential to number and track the trajectories of multiple individuals to ensure the safety and order of children using the facilities.

The DeepSORT algorithm excels in trajectory matching, thus this paper selects the DeepSORT algorithm to perform the trajectory acquisition part of the skeleton sequence. This study uses a YOLOv5-based object detector to identify and locate multiple individuals in the video stream images, and then inputs the human bounding boxes into the DeepSORT-based human tracker to track the movement trajectories of each individual. In this way, the skeleton key point extractor can collect multi-person skeleton sequence data after obtaining continuous tracking information.

1) Optimization of skeleton key point extractor: HRNet (High-Resolution Net) is an efficient feature extraction deep learning network structure specifically designed for key point extraction in human pose estimation tasks. This network adopts a top-down approach, starting from the global perspective of the image and gradually refining to the local key points of each individual, achieving precise human skeleton detection [22]. In the network structure, HRNet utilizes residual modules and up-sampling and down-sampling operations to achieve interaction and fusion between features of different resolutions. By regressing heatmaps to represent the positions of key points and using convolutional networks to extract features and fuse them at multiple scales, HRNet introduces low-resolution features while maintaining the expression capability of high-resolution images, resulting in a more comprehensive and detailed representation of image features through the fusion of different resolution features.

Heatmaps can visually display the prediction of each skeleton key point, where the color intensity represents the confidence of the key point, with darker colors corresponding to higher confidence. This visualization method clearly shows the position of each key point in the image and its corresponding confidence. The process of predicting skeleton key points by regressing heatmaps is illustrated in Fig. 5.

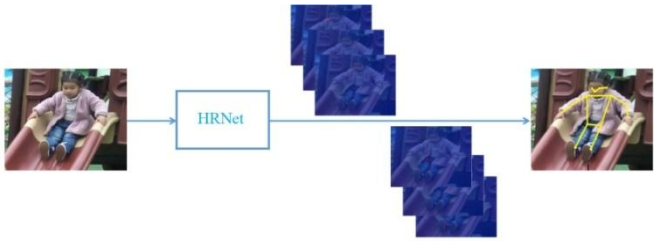


Fig. 5. Regression heatmap prediction diagram (Parental Consent Obtained).

The obtained skeleton key points can serve as fundamental features for deep learning models to analyze human behavior, capturing subtle changes in human posture and spatial relationships. This paper employs an improved human skeleton key point extraction network that incorporates the design philosophy of HRNet, which combines high and low-resolution semantic information while maintaining low-resolution features at higher levels as much as possible. By improving the network's feature fusion module, the recognition accuracy of the skeleton key point extraction network is enhanced, and latency is reduced, thereby improving detection performance.

In the HRNet network, the method of fusing high and low-level features is a crucial part of determining network performance. However, there are issues with this fusion method, particularly the introduction of information errors in some high-

resolution feature maps. The root cause lies in the bilinear interpolation operation used during fusion, which disrupts the symmetry of image pixels and causes pixel shifts, leading to distorted information in the feature maps. This paper proposes an improvement by introducing a Flow Alignment Module (FAM), inspired by the FlowNet algorithm [23]. This algorithm is primarily used to capture optical flow information between adjacent video frames. By incorporating the Flow Alignment Module into the HRNet network, information errors occurring during the fusion of high and low-level features can be effectively resolved. This module adaptively adjusts the alignment between high and low-level features based on the semantic information of the current feature map, reducing the impact of pixel shifts and maintaining pixel symmetry as much as possible. This improvement ensures better consistency of semantic information during network feature fusion, enhancing network performance and effectiveness. The flow alignment module is illustrated in Fig. 6.

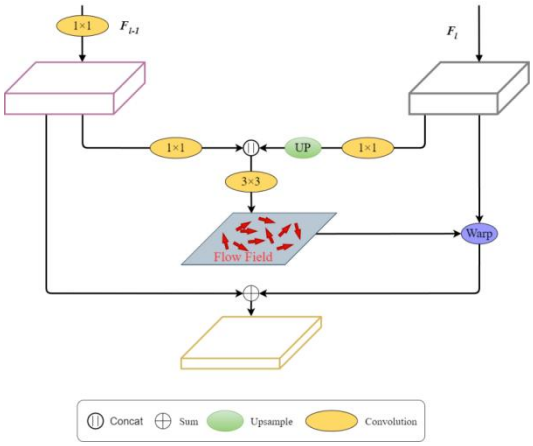


Fig. 6. Flow alignment module.

Among them, F_l represents the high-level low-resolution feature map, and F_{l-1} represents the low-level high-resolution feature map. The function of this module is to ensure that the image size and resolution of adjacent levels match during feature fusion. First, F_l is up-sampled using bilinear interpolation to obtain the same size as F_{l-1}. Then, a 3×3 depthwise separable convolution is used to generate the semantic flow field Δ_{l-1} and perform the flow alignment operation.

For the bilinear interpolation up-sampling method, the pixel point $p_{l-1}(x_{l-1}, y_{l-1})$ in the low-level feature map is mapped to the pixel point $p_l(x_l, y_l)$ in the high-level feature map. Interpolation is performed on the four neighboring points of P_l, where w_l and h_l are the width and height of F_l, and w_{l-1} and h_{l-1} are the width and height of F_{l-1}.

$$(x_l, y_l) = \left(x_{l-1} * \frac{w_l}{w_{l-1}}, y_{l-1} * \frac{h_l}{h_{l-1}}\right) \quad (1)$$

For the flow alignment operation based on the semantic flow field, for each pixel point P_{l-1} in the high-resolution low-level feature map, the following formula is used:

$$p_l = \frac{p_{l-1} + \Delta_{l-1}(p_{l-1})}{2} \quad (2)$$

The pixel point P_l in the low-resolution high-level feature map is obtained through the mapping, and then interpolation is performed on the four neighboring points of P_l. This ensures that the sizes of adjacent low-level feature maps remain consistent. The role of the semantic flow field is to guarantee that under the condition of having a broad field, a high-resolution image with richer semantic information is obtained.

The multi-scale attention mechanism involves inputting features of multiple scales into an attention module or combining multi-scale feature contexts within a single attention module to achieve more comprehensive information extraction. The former approach aggregates feature contexts with consistent scales, effectively capturing and utilizing both low-level detail features and high-level semantic features. The latter approach, also known as multi-scale spatial attention, aggregates feature contexts using convolutional kernels of different sizes or pyramid structures within the attention module. Feature fusion is typically achieved through simple linear operations such as summation and concatenation. This method not only reduces the computational speed of the human skeleton key point extraction algorithm but also decreases the extraction accuracy.

To address the aforementioned issues, this study introduces an Attention-aware Feature Fusion (AFF) module at the output stage of the HRNet algorithm. The multi-scale channel attention module (MS-CAM) within AFF is designed to more efficiently fuse feature information at different scales. This module follows the ideas of ParseNet [24], combining local and global features in CNN neural networks as well as spatial attention and multi-scale feature context aggregation within the attention module. The MS-CAM module can adjust the scale of spatial pooling to control the attention weights in multi-scale feature fusion, enhancing the model's ability to capture semantic information. Additionally, it can combine local and global contexts to maintain the model's lightweight nature and efficiency.

For local channel context aggregation, pointwise convolution (PWConv) is used as a parameter-efficient method. This method utilizes local channel interactions at each spatial position, effectively reducing the model's parameter count while maintaining its performance. The bottleneck structure calculates the local channel context $L(X) \in \mathbb{R}^{C\times H\times W}$:

$$L(X) = B\left(PWConv_2\left(\delta\left(B(PWConv_1(X))\right)\right)\right) \quad (3)$$

The kernel sizes of PWConv1 and PWConv2 are C/r×C×1×1 and C×C/r×1×1, respectively, where L(X) has the same shape as the input features and can retain and highlight fine details in the low-level features. Given the global channel context g(X) and the local channel context L(X), the refined feature $X' \in \mathbb{R}^{C\times H\times W}$ is obtained through the MS-CAM module using the following formula.

$$X' = X \otimes M(X) = X \otimes \sigma\left(L(X) \oplus g(X)\right) \quad (4)$$

In the formula, $M(X) \in \mathbb{R}^{C\times H\times W}$ represents the attention weights generated by the MS-CAM module, $\oplus$ denotes pixel-wise addition, and $\otimes$ denotes element-wise multiplication.

The schematic diagram of the MS-CAM module is shown in Fig. 7.

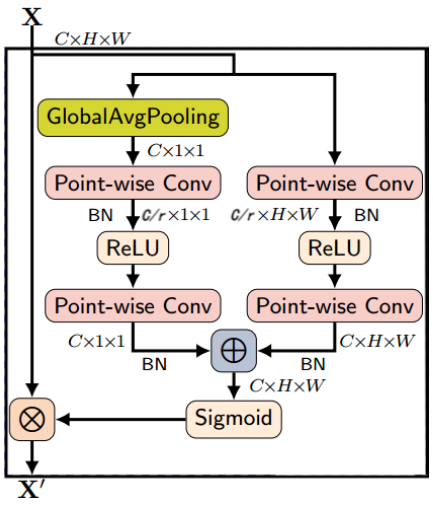


Fig. 7. Multi-Scale channel attention module (MS-CAM).

The two feature maps $X, Y \in \mathbb{R}^{C\times H\times W}$ are input separately, where X and Y are feature maps with different resolutions. According to the Multi-Scale Channel Attention Module (MS-CAM), the Attention-aware Feature Fusion (AFF) is represented by the following formula:

$$Z = M(X \oplus Y) \otimes X + \left(1 - M(X \oplus Y)\right) \otimes Y \quad (5)$$

In the formula, $Z \in \mathbb{R}^{C\times H\times W}$ represents the fused features of X and Y, where $\oplus$ denotes the initial feature integration and element-wise summation as the initial integration. The fusion weights $M(X \oplus Y)$ consist of real numbers, enabling the network to perform weighted averaging between X and Y. The AFF module is illustrated in Fig. 8.

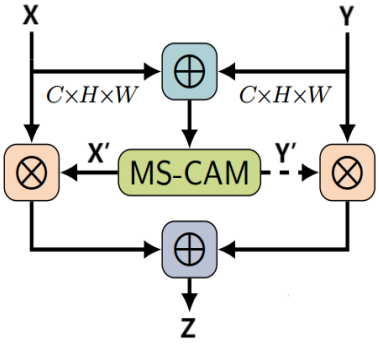


Fig. 8. Attention-aware feature fusion module (AFF).

The Attention-aware Feature Fusion (AFF) module achieves adaptive feature fusion by learning attention weights, enabling the appropriate integration of features from different resolutions and scales. This enhances the HRNet network's focus on features in pose estimation tasks, improving the network's feature consistency and stability, and reducing information discrepancies between feature maps. As a result, the ability to perceive spatial relationships and interactions between target skeleton key points is improved, along with detection accuracy and stability. The overall network structure of HRNet after adding the modules is shown in Fig. 9.

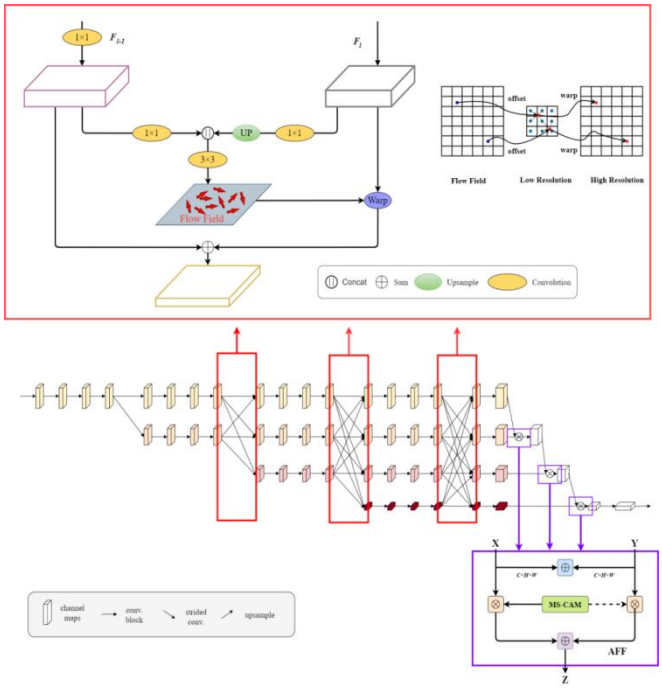


Fig. 9. Improved HRNet network structure (red box indicates flow alignment module, purple box indicates attention-aware feature fusion module, AFF).

Introducing the flow alignment module and attention-aware feature fusion module into the HRNet network aids in parameter computation and algorithmic efficiency, making the network more suitable for real-time detection systems. By aligning image resolutions and adaptively fusing features of different resolutions, these modules can reduce information loss between feature maps and enhance the accuracy of human skeleton key point extraction. This is particularly beneficial in handling complex scenes of slide playground facilities and cases of human occlusion, making the algorithm more applicable.

The skeleton sequence extraction method designed in this paper forms the foundation of the entire algorithm, providing input data for the subsequent human behavior recognition algorithm in slide facilities. The human object detector based on the YOLOv5 network outputs a tensor of dimension N×5, where N is the number of detected people, which will be used in the behavior judgment logic. The number 5 represents the amount of feature data obtained within the bounding box, including the position coordinates of the box (top-left coordinates (x_1, y_1) and bottom-right coordinates (x_2, y_2)), and the confidence score of the box.

After extracting each person's location information, the bounding box information is sequentially fed into the human skeleton key point extraction network to obtain each person's skeleton information. This approach reduces the computational complexity of the extraction network, outputting a tensor of dimension N×V×3, where V represents the predefined number of skeleton key points in the dataset, which is 17, and 3 represents the number of features for each individual key point, including the position coordinates (x_1, y_1) and the confidence score of the key point coordinates.

The DeepSORT algorithm-based human tracker is used for the classification and tracking of each person. The process of collecting skeleton sequences alternates between human tracking and human skeleton key point extraction. Through this process, multiple target tracking trajectories can be obtained and these trajectories can be traversed to collect each person's skeleton data according to the images in the video sequence. Once T frames of skeleton data are collected, these data will form the skeleton sequence information input for the human behavior recognition algorithm, resulting in a tensor of dimension N×C×T×V, where C is the number of features for each individual key point, with a value of 3, and T is the length of the skeleton sequence information. The final tensor will be used as the input data for the human behavior recognition algorithm.

2) Skeleton sequence algorithm and performance experiments: In practical application scenarios, obtaining datasets for hazardous behaviors in slide facilities poses significant challenges. Therefore, data augmentation methods are employed to increase the data volume. By performing operations such as translation, flipping, cropping, rotation, and adding noise to the images, the dataset's content can be enriched, enabling the model to better learn the target features. Additionally, to handle the blank areas that may arise during transformation, black padding is used to reduce the impact on target features. Through these data processing methods, the issue of insufficient data in practical scenarios can be better addressed, thereby enhancing the model's performance and application effectiveness.

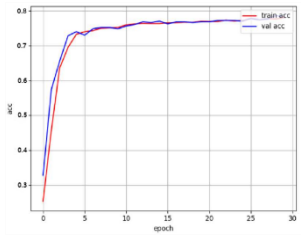
To verify the advantages of the improved HRNet model, we used the original HRNet network, the HRNet network with the flow alignment module, and the HRNet network with both the flow alignment module and the attention-aware feature fusion module as the networks for extracting skeleton key point features. By reasonably designing the module parameters to enhance computational effectiveness, we aimed to achieve good performance while minimizing the increase in computational load, thus obtaining good computational efficiency to better adapt to real-time detection. Ablation experiments on human skeleton key point extraction were conducted on a self-built dataset, with all three groups set to 10 training rounds. The experimental results are shown in Table II.

TABLE II ABLATION EXPERIMENT RESULTS

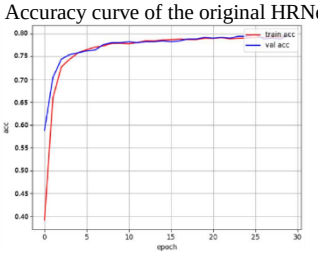
Model	FPS	Computation (G)	mAP(%)
HRNet	51	18.2	75.5
HRNet+FAM	56	19.7	77.8
HRNet+FAM+AFF	58	21.1	79.1

The table data shows that after adding the flow alignment module, the network's mean average precision (mAP) increased by 2.3%, but the computation increased by 1.5G. This indicates that with a slight increase in computation, the network accuracy was improved. Based on this improvement, the attention-aware feature fusion module was further introduced, resulting in an additional increase of 1.4G in computation, while the mAP increased by another 1.3%. Compared to the initial network, although the computational complexity was slightly increased, the network's accuracy, frame processing rate, and computation rate were significantly improved.

The improved algorithm was used to train the human skeleton key point recognition model, and the training results were compared with those of the original HRNet network. The accuracy curves are shown in Fig. 10.



(a) Accuracy curve of the original HRNet



(b) Accuracy curve of the improved HRNet

Fig. 10. Comparison of model training accuracy curves.

As shown in Fig. 10, the model stabilizes after 20 training epochs both before and after the HRNet network improvements. The Fig also indicates that the training model converges faster and achieves better results after the feature fusion module improvements. Finally, a comparison of model accuracy after 30 training epochs is presented in Table III.

TABLE III ACCURACY COMPARISON AFTER 30 TRAINING EPOCHS

Model	Train Acc(%)	Val Acc(%)
HRNet	76.1	75.8
HRNet+FAM+AFF	79.5	79.3

Experimental results indicate that the improvements made to the model significantly enhanced the accuracy of pose recognition and accelerated the computation speed. Therefore, this model is suitable for use as the pose estimation network in the real-time behavior detection algorithm.

D. Design of Human Hazardous Behavior Recognition Algorithm

This study first conducted a survey on the types of slide facility accidents domestically and internationally, and established a checklist of accident types and behaviors. Based on this, the hazardous behavior issues to be addressed in this study were determined. Next, by identifying the recognition categories, a parameterized model of behavior actions was constructed. An algorithm for behavior recognition was designed based on the information obtained from the improved skeleton sequence extraction network mentioned earlier. Finally, behavior recognition was performed using both pose estimation algorithms and sensor-acquired information, and comparative experiments were conducted to verify the feasibility of the algorithm.

In cases of slide playground accidents, improper behavior by users and inadequate supervision by managers can be predicted through extensive observation, statistics, and analysis [25]. The causes of human errors are complex. Based on relevant facility standards and the investigation of slide playground facilities, accidents on slides were categorized and summarized. The design of this detection algorithm needs to complete the identification of behavior categories including crowding, close proximity, orientation abnormality, climbing, staying still, falling, and normal state.

3) Behavior feature analysis and skeleton key point selection: Crowding and close proximity scenarios share a common characteristic: the presence of multiple people in the facility, distinguishing them from other hazardous scenarios. These hazardous behaviors are relatively easy to identify and can be prevented by limiting the number of users. Past object detection algorithms have already obtained the number of human image frames when extracting human information from images, so the number of people can be used as a priori condition for behavior judgment. For distance judgment, the intersection over union (IoU) between each human image frame is used; if the IOU exceeds a preset threshold, it is determined to be too close.

When using the slide, orientation abnormalities occur when people slide down in a non-seated position, meaning the upper body is positioned below the lower body during the slide, which can easily lead to head injuries. To address this issue, the height difference of body key points can be used as a basis for judgment. Since the nose and ankle positions are less likely to be occluded, their detection is relatively stable. Therefore, by comparing the vertical coordinate heights between the middle point of the left and right ankles and the middle point of the nose, the correctness of the sliding orientation can be determined.

During sliding on the slide, the vertical acceleration value is usually less than the gravitational acceleration ggg. When the body falls off the slide, its acceleration value should be equal to the gravitational acceleration. Thus, the midpoint between the shoulders and hips can be used as the body center point. When the vertical acceleration of this center point approaches g, it can be judged that the body has detached from the slide.

Position changes during slide use can be categorized into three situations: climbing, staying still, and normal sliding. During climbing, the posture is not fixed, but the general trend is climbing from bottom to top; staying still refers to the body actively or passively remaining stationary at any position on the slide; and normal sliding refers to sliding from top to bottom without hazardous behaviors. Because small changes in position significantly affect the judgment, the stable and information-rich body center point continues to be used as the judgment basis, with its vertical coordinate changes proving the position change of the body.

Using the improved HRNet algorithm, data for 17 human skeleton key points are obtained. The schematic diagram of the predicted skeleton key point positions is shown in Fig 11, and the correspondence between the skeleton key point names and feature point numbers is shown in Table 4.

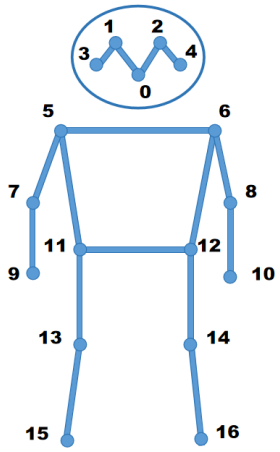


Fig. 11. Schematic diagram of the predicted positions of 17 skeleton key points.

TABLE IV CORRESPONDENCE BETWEEN KEY POINT NAMES AND FEATURE POINT NUMBERS

Feature Point Number	Key Point Name	Feature Point Number	Key Point Name
0	Nose	9	Left Wrist
1	Left Eye	10	Right Wrist
2	Right Eye	11	Left Hip
3	Left Ear	12	Right Hip
4	Right Ear	13	Left Knee
5	Left Shoulder	14	Right Knee
6	Right Shoulder	15	Left Ankle
7	Left Elbow	16	Right Ankle
8	Right Elbow		

The extracted human key points include the nose, eyes, left and right wrists, and left and right ankles, represented by (x, y) coordinates to indicate the positions of each key point. Taking the information [8.33e2 9.26e2 9.44529235e-1 9] as an example to explain the content, 9 represents the 9th joint, which is the left wrist; 9.44529235e-1 indicates the confidence level of detecting this joint; 9.26e2 represents the vertical coordinate pixel value of the key point; and 8.33e2 represents the horizontal coordinate pixel value of the key point. The representation method for other key points is the same.

To achieve behavior recognition functionality, logical settings are applied to the obtained human key point coordinates. In the practical application scenarios of this study, key points 0, 5, 6, 15, and 16 are used to express the parameterized design of several behaviors.

4) Parameterized representation of behaviors: The assessment of crowding and close proximity behaviors relies on human information in the images. During this process, the number of people in the image can be obtained using the number of human bounding boxes extracted by the previous object detection algorithm. Behaviors involving distance issues include children sliding on an adult's lap and pushing on the slide. Due to occlusions, using skeleton key points for distance judgment is somewhat difficult. Therefore, in this study, the intersection over union (IOU) of human bounding boxes is chosen as the basis for distance judgment. When the IOU

reaches a certain threshold, it is determined that the distance between the two is too close.

The upper-left pixel coordinates of the first person's bounding box are (x_{a1}, y_{a1}), and the lower-right pixel coordinates are (x_{a2}, y_{a2}). The upper-left pixel coordinates of another person's bounding box are (x_{b1}, y_{b1}), and the lower-right pixel coordinates are (x_{b2}, y_{b2}). The area of bounding box A is $S_A = (x_{a1} - x_{a2}) \times (y_{a1} - y_{a2})$, and the area of bounding box B is $S_B = (x_{b1} - x_{b2}) \times (y_{b1} - y_{b2})$. The IOU schematic diagram is shown in Fig 12.

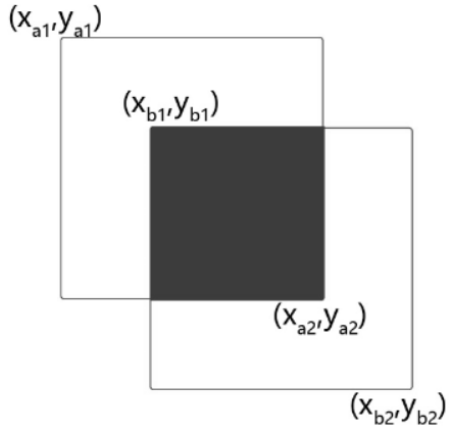


Fig. 12. IOU schematic diagram.

To address the issue of slide orientation, the coordinates of the detected human nose key point (0) and the midpoint coordinates of the ankle joints (15, 16) are used for judgment. The height difference between the nose and ankles is used to make the determination. In a stable state, the nose height is denoted as $\overline{H_{nose}}$, the ankle height is denoted as $\overline{H_{ankle}}$, and the average height of both ankles is denoted as $\overline{H_{ankles}} = (H_{lankle} + H_{rankle})/2$. The difference between the nose height and the average height of both ankles in a stable state is recorded as $D_{na} = \overline{H_{nose}} - \overline{H_{ankles}}$. If the height difference is negative, it indicates an orientation abnormality.

For the determination of falling, climbing, staying still, and normal states, the average value of the coordinates of the left shoulder (5), right shoulder (6), left hip (11), and right hip (12) key points is used as the coordinate value of the human center point. Specifically, $\overline{x_{center}} = (x_{lshoulder} + x_{rshoulder} + x_{lhip} + x_{rhip})/4$, and $\overline{y_{center}} = (y_{lshoulder} + y_{rshoulder} + y_{lhip} + y_{rhip})/4$. If $\overline{y_{center}}$ continues to increase, it indicates that the person is climbing rather than sliding. Conversely, if it decreases, it indicates normal sliding. If it remains unchanged, it indicates that the person is staying still.

To determine falling behavior, the center point's coordinates are first calculated. Then, the second derivative of the pixel coordinates is computed to obtain the center point's acceleration. The expressions for calculating the vertical velocity and acceleration are as follows:

$$v_{ij} = \frac{(y_{ij} - y_{i-1,j})}{\Delta t} \quad (5)$$

where y_{ij} is the vertical coordinate of the j-th key point in the i-th frame, $y_{i-1,j}$ is the vertical coordinate of the j-th key point in the $(i-1)$-th frame, and Δt is the time interval between two adjacent frames.

$$a_{ij} = \frac{(v_{ij} - v_{i-1,j})}{\Delta t} \quad (6)$$

where v_{ij} is the velocity of the j-th key point in the i-th frame, and $v_{i-1,j}$ is the velocity of the j-th key point in the $(i-1)$-th frame.

The algorithm designed in this study follows these steps: first, load the weight file and initialize the recognition model, then recognize humans and slides. Based on the detected number of people, the algorithm proceeds as follows: if the number of people is 2, further calculate the intersection over union (IoU) and obtain information on distance, climbing status, etc.; if the number of people is 3 or more, output "crowded"; otherwise, perform person comparison. During the recognition process, the algorithm also detects human landmarks, acceleration, and other parameters to determine if there are any abnormal conditions, and finally sends the results to the monitoring interface for real-time surveillance.

Based on the above definitions and analyses of various behaviors, combined with the investigation and practical experience of slide accidents, a series of behavioral characteristic indicators were designed to achieve targeted monitoring of the usage behavior of slide playground facilities. Subsequently, experimental methods were used to verify the reliability of the proposed behavioral indicators. The behavioral characteristic indicators are shown in Table 5.

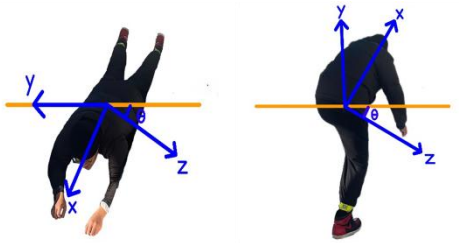
TABLE V PARAMETERS OF HUMAN BEHAVIOR CHARACTERISTICS

Behavior type	Feature parameter	Parameter indicator
Crowding	Number of human bounding boxes	Num≥3
Close proximity	Number of human bounding boxes IOU value	Num=2,IOU＞0.3
Orientation abnormality	Nose key point vertical coordinate Ankle key point vertical coordinates	D_{na}＜0
Falling	Vertical acceleration value of the human center point	9.5m/s^2≤a_{ij}≤9.8m/s^2
Climbing	Vertical coordinate of the human center point	Continuously increasing vertical coordinate
Staying still	Vertical coordinate of the human center point	Unchanging vertical coordinate
Normal state	Vertical coordinate of the human center point	No other category present Continuously decreasing vertical coordinate

5) Recognition logic design: In the establishment of the database, the selection of action materials must follow certain strategies[26][27][28]. The categories of behavior recognition through pose information in this study include five types: orientation abnormality, climbing, staying still, falling, and

normal state. The recognition logic design is achieved by setting changes in pose angles, the direction of acceleration, and acceleration value thresholds.

The judgment of two behaviors, orientation and climbing on the slide, can be made based on the positive or negative values of the X-direction acceleration a_x obtained from the sensor. These values represent whether the body is oriented upward or downward in that direction. Additionally, the Z-direction pose angle $AngleZ$, which is typically described as the pitch angle and usually denoted by θ, describes the angle between the body's front orientation and the horizontal plane, as shown in Fig 13. When the body orientation is abnormal, θ is a negative acute angle, and a_x is positive. When the body is climbing the slide, θ remains a negative acute angle, but a_x is negative. Based on this information, these two behaviors can be identified.



(a) Falling pose (b) Climbing pose

Fig. 13. Pitch angle pose information.

To address the issue of distinguishing between different sliding states on the slide, staying still refers to the condition where position information does not change. This can be determined by using the three-axis velocity. The speed information is obtained by integrating the acceleration data, and the calculation expression is:

$$v_i(t) = \int_0^t a_i(t)dt \quad (7)$$

where iii represents the three-axis directions, with the velocities in the x, y, and z directions calculated separately. When the changes in v_x, v_y, and v_z are minimal and nearly zero, staying still can be determined. As previously described, the falling condition is indicated by the vertical acceleration being equal to the gravitational acceleration. This condition is met only in this behavior, resulting in a low probability of misjudgment. The normal sliding state should meet the condition that θ is always positive. The seated position on the slide may involve the body being close to the slide or sliding upright, so both acute and obtuse angles are possible. Additionally, the vertical velocity should always be downward. If both conditions are met, normal sliding can be determined.

6) Experimental results validation: Using the skeleton sequence extraction algorithm based on the improved HRNet, a human motion dataset was created, including seven types of action postures: crowding, close proximity, orientation abnormality, climbing, staying still, falling, and normal state. The data collected by the camera is input into the skeleton extraction network to extract human pose information from the images. In this process, each person's pose is recognized and represented as a dataset composed of the two-dimensional coordinates of 17 key points. These key point information is saved as train.txt text files and organized into train.csv files through scripts. Additionally, normalization is performed to eliminate the impact of dimensional differences between data. The dataset includes a total of 4200 skeleton sequence images, with a training set to validation set ratio of 8:2. The sample distribution of the human behavior dataset for the slide facilities in this study is shown in Table VI.

TABLE VI KEY POINT NAMES AND FEATURE POINT NUMBERS CORRESPONDENCE

Behavior category	**Crowding**	**Close Proximity**	**Orientation Abnormality**	**Climbing**	**Staying Still**	**Falling**	**Normal State**
Skeleton Sequences (segments)	605	510	460	640	520	560	805

To verify the effectiveness of the behavior recognition algorithm proposed in this paper, experiments were conducted using the improved HRNet network on a self-built skeleton sequence dataset. A total of four testers participated in the experiments, each performing the aforementioned seven types of behaviors for data collection. The experimental data results are shown in Table VII.

TABLE VII EXPERIMENTAL DATA RESULTS

Behavior	**Number of tests**				**Accuracy /%**
	1	2	3	4	
Crowding	121	134	169	146	94.2
Close Proximity	105	127	144	118	96.9
Climbing	143	167	133	140	91.1
Staying Still	125	113	142	108	93.8
Orientation Abnormality	107	114	109	115	96.7
Falling	133	151	173	102	99.8
Normal State	203	187	223	181	98.6

The experimental results show that the human behavior detection algorithm based on the improved HRNet network design achieved an average detection accuracy of over 90%. It performed exceptionally well in recognizing behaviors such as normal state and falling, as these behaviors have relatively distinct features. However, there are challenges in recognizing behaviors such as climbing and staying still, mainly due to the key points in the judgment logic being affected by trunk occlusion or the complexity of the actions, leading to unstable recognition and resulting in misjudgments.

The posture sensor used in this study is the WT901WIFI, an integrated 9-axis motion analysis component that combines a high-precision gyroscope, accelerometer, and geomagnetic sensor. By solving the attitude matrix of the posture calculation system, converting the coordinates of specific forces, and

updating the attitude matrix, it outputs acceleration data, which is integrated over time to obtain the instantaneous velocity of the carrier [29]. When the sensor rotates with the human body, the gyroscope can detect the rotational angular velocity of the carrier. To obtain the human body's motion posture information, the angular velocity output by the gyroscope needs to be integrated, which provides the angular increment relative to the reference coordinate system, thus deriving the motion posture information. The smaller the time increment of integration, the higher the accuracy of the obtained angular data. After acquiring the attitude angle data, the human body's positional information can be derived through secondary integration, thereby obtaining the position change in three-dimensional space[30]. The internal integration of the attitude solver and dynamic Kalman filtering algorithm within the device allows the sensor to accurately output posture in dynamic environments, facilitating behavior recognition based on angle information.

In the field of human action recognition based on accelerometers, many studies have detailed the data collection process [31][32][33]. Using self-collected acceleration data to train and test recognition algorithms, the recognition rate largely depends on the quality of the collected database. To enhance the comparability of this experiment's results, a dataset based on a nine-axis accelerometer was established. Before data collection, the sensor needs to be fixed, ensuring the three-axis directions measured by the sensor completely coincide with the three-axis directions of the measured equipment to ensure data reliability. Additionally, the sensor must be firmly fixed to prevent shaking, which could cause significant measurement errors in acceleration data. Lin[31] and colleagues collected posture information by placing sensors at different wearing positions, including the waist, wrist, ankle, arm, and thigh, comparing the impact of each position on recognizing daily motion patterns. Results showed that the wearing position significantly affected posture recognition rates, with the highest recognition rate achieved when the sensor was fixed at the waist. Therefore, in this collection, the sensor was fixed at the waist to test the acquisition of posture information. During the study, multiple data collections were conducted, with field equipment collecting data on the slide facilities of a kindergarten. The collected data included three-axis acceleration, three-axis angular velocity and angle, and the corresponding collection time.

The algorithms based on machine vision information and those based on pose sensor information were tested on the constructed dataset. The accuracy confusion matrices for recognizing human behaviors in slide facilities for both approaches are shown in Tables VIII and IX.

The detection methods based on machine vision and sensors both showed good performance in terms of recognition accuracy, thereby validating the generalization capability of the algorithm established in this study for recognizing human behaviors in slide facilities. According to the data in Table VIII, the recognition rates of different behaviors in the pose estimation scheme show certain differences. The root cause of this difference can be traced to the stability and latency of obtaining skeleton sequences during real-time detection. When the key point information used by the algorithm is not updated in time during the judgment process, it may lead to misjudgments or omissions. Therefore, setting parameters between frames is crucial for behavior recognition in practical situations.

TABLE VIII CONFUSION MATRIX FOR BEHAVIOR RECOGNITION BASED ON MACHINE VISION INFORMATION

	Climbing	**Staying Still**	**Orientation Abnormality**	**Falling**	**Normal State**
Climbing	0.9013	0	0	0	0
Staying Still	0.0655	0.9537	0.0046	0	0.0233
Orientation Abnormality	0	0	0.9735	0	0.0014
Falling	0	0	0	1	0
Normal State	0.0332	0.0463	0.0219	0	0.9753

TABLE IX CONFUSION MATRIX FOR BEHAVIOR RECOGNITION BASED ON SENSOR INFORMATION

	Climbing	**Staying Still**	**Orientation Abnormality**	**Falling**	**Normal State**
Climbing	0.9115	0.0086	0.0281	0	0
Staying Still	0.0742	0.9383	0.0046	0	0.0134
Orientation Abnormality	0	0	0.9673	0.0004	0
Falling	0	0	0	0.9985	0
Normal State	0.0143	0.0531	0	0.0011	0.9866

The recognition scheme based on pose information also has its advantages and disadvantages. By comparing and analyzing the experimental results, it can be seen that the recognition accuracy of the two schemes is shown in Fig. 14.

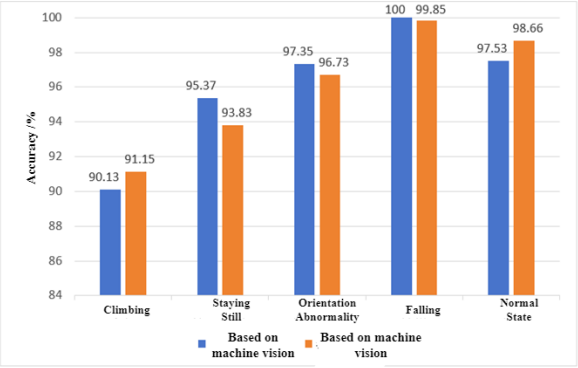


Fig. 14. Comparison of recognition accuracy of the two methods.

Both methods perform poorly in recognizing climbing behavior. This is mainly due to the complexity and inconsistency of climbing actions, as well as possible occlusion. Therefore, there are significant challenges in obtaining key points and setting pose information thresholds. Recognition of body orientation shows that the method based on skeleton sequences is superior to the judgment based on pose angles. Climbing behavior is not limited to the user lying face down on

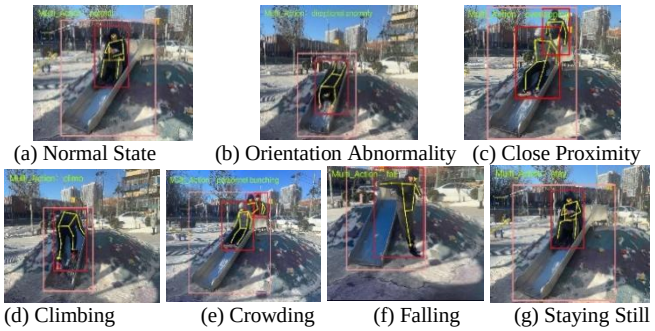
the slide but can also include sliding with the back against the slide and head down, which causes the pitch angle judgment to fail. However, the judgment based on skeleton key points is relatively stable during detection. Although there may be occasional misjudgments due to interrupted behavior, the overall recognition accuracy is higher.

For recognizing staying still and normal sliding behaviors, both methods show generally stable performance, with accuracy rates of over 96%. Misjudgments are mainly related to the set judgment time, as delays in camera capture and sensor transmission can affect the stability of frame-to-frame information. The accuracy of determining falling is nearly error-free because this behavior has obvious characteristics, and obtaining acceleration value information is relatively easy, making the processing methods more diverse and less prone to errors.

By comparing the recognition accuracy of human behaviors in some slide facilities, the results show that the method based on human skeleton sequence information performs well in recognizing the above behaviors. It effectively reduces the impact of complex human postures and inconsistent behavior scenarios.

III. Experimental Results

The slide facility human behavior monitoring system is based on the human behavior recognition algorithm presented in this paper and is deployed on a computer upper platform. The computer uses an external camera to capture images of the monitored scene and stores them in real time. The monitoring system invokes the human behavior recognition algorithm to judge various hazardous behaviors from the images and presents the processed results on the system's interactive interface, while also controlling the computer's buzzer to sound an alarm. To obtain a comprehensive monitoring view, the external USB camera is fixed at positions 1 meter and 3 meters from the ground and the slide, respectively. The system conducts detection tests on seven types of behaviors, and the detection effects are shown in Fig. 15.



(a) Normal State (b) Orientation Abnormality (c) Close Proximity

(d) Climbing (e) Crowding (f) Falling (g) Staying Still

Fig. 15. Detection results of the behavior recognition system.

To verify the effectiveness of the human behavior recognition based on the PyQt interactive interface system, tests were conducted after the real-time collection of a series of behaviors, with 90 groups recorded for each behavior. The confusion matrix for recognizing seven types of behaviors by the human behavior recognition model tested on a computer upper platform system built with PyQt is shown in Fig. 16.

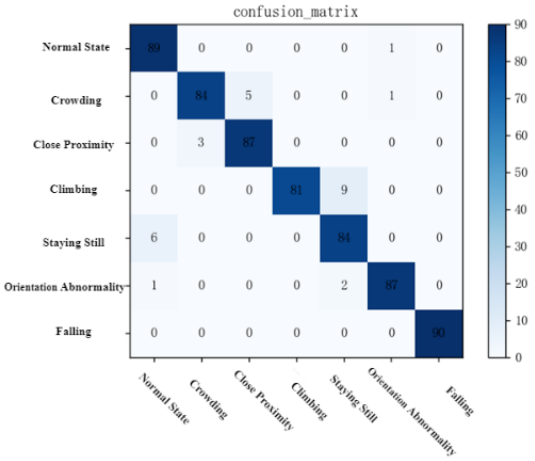


Fig. 16. Confusion matrix of system behavior detection results.

After experimental testing, the test accuracy, recall, specificity, and F_1 score were calculated from the obtained data to serve as the basis for evaluating the system's detection performance. The calculated data are shown in Table X.

TABLE X Distribution of Test Performance

Behavior Category	Precision	Recall	Specificity	F1 Score
Normal State	0.988	0.975	0.993	0.981
Crowding	0.933	0.924	0.941	0.928
Close Proximity	0.966	0.933	0.982	0.949
Climbing	0.9	0.885	0.924	0.892
Staying Still	0.933	0.926	0.949	0.929
Orientation Abnormality	0.966	0.956	0.975	0.961
Falling	1.0	0.987	1.0	0.993

The main diagonal elements of the confusion matrix reflect the number of correctly recognized specific behavior categories. The recognition accuracy is the ratio of correctly classified behaviors to the total number of classified outputs [34]. According to the confusion matrix results, the recognition accuracy for various behaviors is generally high in the upper computer system. However, Table 10 shows that except for the recognition accuracy of climbing behavior, which is 0.9, the accuracy for other behaviors is above 0.9.

Analyzing this, the judgments for crowding and close proximity rely on determining the number of humans after object detection. This process is influenced by logical similarities, which may lead to misjudgments between them. Climbing behavior may involve pauses, making it easy to be misjudged as staying still. Additionally, the posture of climbing behavior may cause limb occlusion, affecting the extraction of human skeleton key points, resulting in less comprehensive skeleton sequence features and relatively lower recognition accuracy.

Overall, the detection data show that the average recognition performance for the seven types of behaviors is satisfactory and meets the expected goals of this study.

IV. CONCLUSION

This study delves into the issue of human hazardous behavior recognition in slide facilities, designing a recognition method based on a pose estimation extraction algorithm for human skeleton sequence information and pose parameterization representation algorithm. PyQt technology was used to deploy the detection model on a computer upper platform to recognize hazardous behaviors in slide facilities. The main contributions of this paper are as follows:

1) Proposed an improved method for extracting human skeleton sequences: This method includes detecting humans within the range of the slide and slide recognition box in the scene and tracking human trajectories. The improved HRNet network is used to extract continuous skeleton sequence data for each person.

2) Proposed improvement solutions for issues within the HRNet network: The solutions address problems of pixel symmetry being disrupted during the process of obtaining high-resolution features from low-resolution feature maps and the loss of features during the fusion of different resolutions. A network model incorporating a flow alignment module (FAM) and an attention-aware feature fusion module (AFF) was proposed. Experimental results show that the network integrating these two modules, compared to using only the HRNet network and the HRNet network with the flow alignment module, improves accuracy on a self-built dataset. The accuracy of hazardous behavior detection increased by 3.6% with a slight increase in training complexity, achieving good computational efficiency and accuracy.

3) Designed a human hazardous behavior recognition algorithm for slide facilities: By organizing a list of hazardous behaviors in the scene, summarizing hazardous behaviors on slides, and conducting parameterized pose design. The skeleton key point sequence information of humans sliding extracted by the improved HRNet network and DeepSORT tracking network is combined with the image classification information obtained by the object detection network, and input into the parameterized pose representation algorithm to determine the behavior category of the users' poses.

The human behavior monitoring system for slide facilities designed in this study achieves non-contact equipment safety management through machine vision technology. This method avoids the impact of the equipment on children during the sliding process and helps improve the digitalization, informatization, and intelligence levels of reasonable supervision of amusement equipment in places like playgrounds, schools, and communities. Nevertheless, this study has certain limitations, as not all mentioned technologies were deeply explored. Future work should aim to further improve:

1) Segmenting and supplementing behavior categories: Currently, only parameterized design and experimental verification for hazardous behaviors in investigated safety accidents have been conducted. In the future, behavior categories can be further segmented and supplemented, collecting more human behavior information in slide facilities, designing relevant parameter expressions, and further enhancing the model's applicability in slide facility scenarios.

2) Enriching image feature information: The current human behavior detection methods utilize relatively single image feature information, and pose parameterization design should not be limited to information such as acceleration, position, and angle. Future research will use facial recognition technology to achieve expression detection of human targets to assist in human behavior detection.

3) Introducing three-dimensional image information: The slide scene images and videos collected in this study are all two-dimensional, lacking the reliability of three-dimensional spatial information. Therefore, future research will consider using binocular vision cameras to collect three-dimensional images, employing three-dimensional reconstruction technology and reasonably designing the representation of human spatiotemporal action information.

REFERENCES

[1] Hao, Jianfeng. (2009). *Design and Research of Children's Playground Equipment* [D]. Hubei: Hubei University of Technology. DOI: 10.7666/d.Y1551764.

[2] Meng, Lingjun, Yang, Xinming, Fu, Ganwei, et al. (2022). Safety Assessment of In-Use Large Amusement Facilities (Slide Type). *Special Equipment Safety Technology*, 2022(6), 51-53. DOI: 10.3969/j.issn.1674-1390.2022.06.020.

[3] Hayhurst, R Emery. Industrial accident prevention, a scientific approach [J]. American Journal of Public Health and the Nations Health, 1932, 22(1):119-120.

[4] Cui, Wenxiang, Xu, Yanli. (2007). The Relationship Between Preschool Children's Cognition of Accidental Injuries and Accident-Prone Behaviors. *Maternal and Child Health Care of China*, 22(22), 3094-3096. DOI: 10.3969/j.issn.1001-4411.2007.22.026.

[5] Lu, Lei, Xu, Biao, Lin, Shuang, Zhang, Xianliang, & Ge, Wanlei. (2021, November 10). High Strength and Toughness Children's Slide.

[6] Guo H, Yu Y, Ding Q, et al. Image-and-skeleton-based Parameterized Approach to Real-time Identification of Construction Workers'Unsafe Behaviors[J]. Journal of Construction Engineering and Management, 2018, 144(6):04018042.

[7] Yang, Bin, Xiao, Yun, Dong, Kaiwen, Liu, Xixiang, & Huang, Han. (2021). Human's Dangerous Action Recognition in Petrochemical Scene Using Machine Vision. *Laser & Optoelectronics Progress*, 58(22), 3914. doi: 10.3788/LOP202158.2215001.

[8] Wang, Hong, Deng, Yuanshi, Chang, Zhengwei, et al. (2022). Behavior Recognition Technology of Power Workers Based on Deep Learning. *Sichuan Electric Power Technology*, 45(3), 23-28. DOI: 10.16527/j.issn.1003-6954.20220304.

[9] Zhang Y, Ding K, Hui J, et al. Skeleton-RGB integrated highly similar human action prediction in human－robot collaborative assembly[J]. Robotics and Computer-Integrated Manufacturing, 2024, 86: 102659.

[10] Han S, Lee S. A Vision-based Motion Capture and Recognition Framework for Behavior-based Safety Management[J]. Automation in Construction, 2013, 35:131-141.

[11] XIONG Ruoxin, SONG Yuanbin, WANG Yuxuan, DUAN Yanjuan. Application of convolutional neural network-based 3D posture estimation in behavioral analysis of construction workers[J]. China Safety Science Journal, 2019, 29(7): 64-69.

[12] Na-na Fu, Da-ming Liu, Xiao-ting Cheng, et al. Fall detection algorithm based on lightweight OpenPose model. Sensor and Microsystem, 2021, 40(11): 131-134, 138. DOI:10.13873/J.1000-9787(2021)11-0131-04.

[13] Thakkar K, Narayanan P J. Part-based graph convolutional network for action recognition[J]. arXiv preprint arXiv:1809.04983, 2018.

[14] Huang L, Huang Y, Ouyang W, et al. Part-level graph convolutional network for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11045-11052.

[15] Qiu H, Hou B. Multi-grained clip focus for skeleton-based action recognition[J]. Pattern Recognition, 2024, 148: 110188.

[16] Wu L, Zhang C, Zou Y. SpatioTemporal focus for skeleton-based action recognition[J]. Pattern Recognition, 2023, 136: 109231.

[17] Jian-bao Zhu, Zhi-long Xu, Yu-wei Sun, et al. Detection of Dangerous Behaviors in Power Stations Based on OpenPose Multi-person Attitude Recognition. *Automation and Instrumentation*, 2020, 35(2): 47-51.

[18] Qiu, T. H., Wang, L., & Wang, P. (2022). Research on Object Detection Algorithm Based on Improved YOLOv5. *Computer Engineering and Applications*, 58(13), 63-73.

[19] Wang C Y, Liao H Y M, Ye H, I H, Wu, Y H, Chen P Y, Hsieh, J W. CSPNet: A New Backbone that can Enhance Learning Capablity of CNN[C]. In Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, WA, USA, 2020:1571-1580.

[20] Guo K, He C, Yang M, Wang S. A pavement distresses identification method optimized for YOLOv5s[C]. Sci. Rep, 2022:35-42.

[21] CAO Ziqiang, SAI Bin, and LU Xin. Review of pedestrian tracking: Algorithms and applications[J]. Acta Physica Sinica, 2020, 69(8): 084203. doi: 10.7498/aps.69.20191721.

[22] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:5693-5703.

[23] DOSOVITSIY A, FICHER P. ILG E, et al. Flownet: Learning optical flow with convolutional networks[C]. Proceedings of the IEEE international conference on computer vision. 2015:2758-2766.

[24] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better[J]. CoRR, abs/1506.04579, 2015, 12:122-134.

[25] Deng,S.,&Pan, Y. (2022). Fine-grained management of construction workers' unsafe behaviors based on cognitive mechanisms. Journal of Civil Engineering and Management, 39(4), 178-184. https://doi.org/10.13579/j.cnki.2095-0985.2022.20210892

[26] Lin Bao, Intille S S. Activity recognition from user-annotated acceleration data[C]. Proc of the 2nd International Conference on Pervasive Computing. Springer, Berlin, 2004:1-17.

[27] Hull J. A database for handwritten text recognition research[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(5):550-554.

[28] Gao Wen, Cao Bo, and Shan Shiguang, et al. The CAS-PEAL large-scale chinese face database and baseline evaluations[J]. IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans, 2008, 38(1):149-161.

[29] Rahimi Hossein et al. A fast alignment of marine strapdown inertial navigation system based on adaptive unscented Kalman Filter[J]. Transactions of the Institute of Measurement and Control, 2021, 43(4):749-758.

[30] Zhong Yulu, Zhou Zhaihe, Zeng Chuanwei, et al. Quadrotor Attitude Measurement System Design and Implementation Using Quaternion Kalman Filter[J]. Electronic Measurement Technology, 2020, 43(1): 41-45. DOI:10.19651/j.cnki.emt.1903297.

[31] Lin Bao, Intille S S. Activity recognition from user-annotated acceleration data[C]. Proc of the 2nd International Conference on Pervasive Computing. Springer, Berlin, 2004:1-17.

[32] Kern N, Schiele B, and Schmidt A. Multi–sensor activity context detection for wearable computing[C]. In Proc. EUSAI, LNCS, Eindhoven, The Netherlands, November, 2003, 2875:220–232.

[33] Sun Yuhang. Research on Human Motion Pattern Recognition Technology[D].Anhui University of Technology, 2020. DOI:10.27790/d.cnki.gahgy.2020.000258.

[34] Hansong Su, Tengteng Liu, Gaohua Liu, et al. Algorithm for Student Behavior Detection Based on Neural Network. Laser & Optoelectronics Progress, 2020, 57(22): 177-183. DOI:10.3788/LOP57.221016.