Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 2827-2835
Journal homepage: http://iieta.org/journals/ts

An Intelligent Safety Monitoring Approach for Multi-Person Behavior Recognition in Elderly-Child Shared Spaces Based on Video Imagery

Zhiqiang Zhang

Wenxin Academy, The Open University of Henan, Zhengzhou 450046, China

Corresponding Author Email: zhangzhiqiang@haou.edu.cn

https://doi.org/10.18280/ts.420533

Received: 3 February 2025
Revised: 20 August 2025
Accepted: 30 August 2025
Available online: 31 October 2025

Keywords:
elderly-child shared space, video imagery, multi-person behavior recognition, safety monitoring

ABSTRACT

With the deepening of population aging and the growing demand for child care, the number of elderly-child shared spaces—such as community centers and private residences—has been steadily increasing. Due to the physiological characteristics of the elderly and young children, safety incidents frequently occur in such environments. Traditional video surveillance systems that rely on manual monitoring are inadequate for providing real-time and precise safety assurance. Although progress has been made in behavior recognition based on video imagery, existing methods show significant limitations when applied to elderly-child shared spaces. Single-person behavior recognition algorithms often suffer from feature confusion in multi-person interaction scenarios. Multi-target tracking methods are unstable in indoor environments with frequent occlusion. Furthermore, abnormal behavior detection models exhibit poor generalization for behaviors specific to elderly and young populations, resulting in high false alarm rates. This paper proposes an intelligent safety monitoring approach for elderly-child shared spaces based on multi-person behavior recognition from video imagery. The aim is to address the above challenges and enhance the intelligence and reliability of safety monitoring in such environments.

1. INTRODUCTION

With the deepening of population aging [1-3] and the growing demand for child care [4, 5], the number of elderly-child shared spaces [6-9], such as community activity centers, family residences, and elderly-child care institutions, continues to increase. In such spaces, elderly people have slow movements and weak reaction capabilities, while young children lack self-protection awareness, making it easy for safety incidents such as falls, collisions, and getting lost to occur, which puts forward urgent demands for real-time and precise safety monitoring. As an important means of safety protection, video surveillance technology has been widely applied [10-13], but traditional monitoring mainly relies on manual watching, making it difficult to cope with complex scenarios involving simultaneous multi-target activities, with strong delay in abnormal behavior recognition, and it is hard to meet the safety protection requirements of elderly-child shared spaces.

Research on safety monitoring methods for elderly-child shared spaces has important practical significance and application value. From the social level, this research can effectively reduce the incidence of safety accidents among the elderly and children, protect the lives of vulnerable groups, reduce the burden of family care, and maintain social harmony and stability. From the technical level, it promotes the deepened application of video image multi-target behavior recognition technology in specific scenarios, facilitates the cross-integration of computer vision and the field of safety protection, and provides a reference for the scenario-based implementation of related technologies.

Existing research has made certain progress in the field of video image behavior recognition, but there are still obvious deficiencies for the specific scenario of elderly-child shared spaces. The single-target behavior recognition algorithms proposed in literature [14, 15] are prone to feature confusion in multi-target interaction scenarios and lack the ability to distinguish behaviors when elderly and children are active simultaneously. The multi-target tracking methods in literature [16, 17] have poor tracking stability in indoor environments with frequent occlusion, leading to interruptions in the continuity of behavior recognition. The abnormal behavior detection models in literature [18-20] rely on large amounts of annotated data and have weak generalization ability in recognizing behaviors specific to elderly and children, with a high false alarm rate.

This paper focuses on safety monitoring methods for elderly-child shared spaces based on video image multi-target behavior recognition. The main research contents include two aspects: first, multi-target behavior recognition in elderly-child shared spaces based on video image feature enhancement, by optimizing the feature extraction network to strengthen the discriminability of behavior features of elderly and children and improve the accuracy and robustness of multi-target behavior recognition in complex scenarios; second, safety monitoring strategies for elderly-child shared spaces based on video image multi-target behavior recognition, which combine the recognition results to

construct a safety risk assessment model, realizing real-time early warning and emergency response linkage for abnormal behaviors. The value of this research lies in providing a complete solution from behavior recognition to safety management for elderly-child shared spaces, enhancing the intelligence level of safety monitoring, and offering technical support for ensuring the safety of elderly and children.

2. MULTI-TARGET BEHAVIOR RECOGNITION IN ELDERLY-CHILD SHARED SPACES BASED ON VIDEO IMAGE FEATURE ENHANCEMENT

In elderly-child shared spaces, elderly people move slowly and children behave actively. The behavioral features of the elderly and children are significantly different, and complex situations such as limb overlapping and scene occlusion are likely to occur when multiple targets are active simultaneously. These lead to difficulties for traditional single-target recognition algorithms to accurately distinguish the behavioral features of different objects, and multi-target tracking is also prone to feature confusion due to occlusion. At the same time, the annotated data of such scenes is often limited, especially lacking behavioral samples of elderly and children at different scales. Existing models tend to overfit on limited data and are difficult to learn robust behavioral patterns. In addition, the specific behaviors of elderly and children require more delicate semantic features to support recognition, while the feature richness extracted by conventional backbone networks is insufficient to meet the demand for precise recognition.

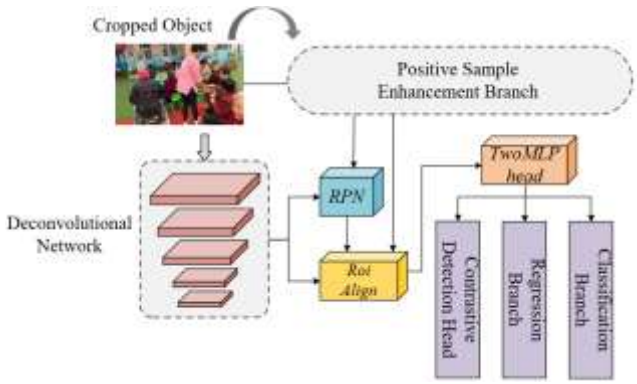

Figure 1. Network structure of a multi-target behavior recognition model for elderly-child shared spaces

To address this, this paper proposes a multi-scale contrastive fine-tuning network for multi-target behavior recognition in elderly-child shared spaces. The model network structure is shown in Figure 1. The specific implementation approach is closely centered on the scene characteristics of elderly-child shared spaces. First, deep deformable convolution is used to optimize the backbone network, enhancing the network's ability to capture behavioral features of different scales and types, and extracting richer semantic features to adapt to the significant differences in movement amplitude and behavior patterns between the elderly and children. Second, object-level features are incorporated during the training stage to supplement the number of positive samples of elderly and child targets at different scales, solving the problem of scarce specific behavior samples in such spaces, enabling the model to learn more robust prior knowledge from the enhanced original features, and improving the stability of identifying specific behaviors of the elderly and children. Third, an improved supervised contrastive branch is introduced to enhance the feature discriminability among different behavior categories, combined with a balanced fine-tuning strategy to balance the training weights of elderly and child behavior samples, ultimately improving classification accuracy on a limited image set, while enhancing the model's anti-overfitting ability and ensuring stable and precise multi-target behavior recognition in real elderly-child shared spaces.

The proposed model is based on the FasterRCNN framework, and layered improvements are made according to the characteristics of elderly-child shared spaces, forming a complete architecture of "feature extraction — region proposal — feature optimization — loss calculation." In the feature extraction layer, the convolution operators of the traditional backbone network are replaced with deformable convolution operators to build an optimized ResNet network. This improvement enhances the network's ability to resist complex interferences in elderly-child shared spaces, such as furniture occlusion and lighting changes, while improving few-shot classification performance, adapting to the characteristics of scarce and significantly different behavior samples of elderly and children in this scene, and providing more robust base features for subsequent recognition. In the region proposal layer, an auxiliary positive sample enhancement branch module is integrated into the region proposal network structure to adaptively capture the multi-scale object-level features of elderly and children, supplementing the number of positive samples at different scales, enabling the model to learn multidimensional prior knowledge of elderly and child behaviors from the base class during the basic training stage and improving the accuracy of candidate region generation. In the feature optimization and loss calculation layer, the model encodes candidate features through an improved spatial pyramid pooling network and incorporates supervised contrastive learning ideas to improve the loss function. Aiming at the characteristics of small intra-class variation and significant inter-class differences in behaviors in elderly-child shared spaces, the loss function strengthens instance-level intra-class compactness and inter-class separability by measuring the similarity between target encodings, allowing the model to more accurately distinguish various behaviors of elderly and children. In the training phase, an improved balanced fine-tuning method is introduced, which dynamically adjusts the training weights of elderly and child behavior samples to alleviate the problem of sample imbalance that may occur in few-shot scenarios, ensuring that the model can stably learn the behavioral features of the two groups in a limited dataset.

2.1 Optimized ResNet network

The core motivation for constructing the optimized ResNet network is to address the adaptation deficiencies of traditional ResNet in multi-target behavior recognition within elderly-child shared spaces. In such spaces, behavior samples are scarce and scenes are complex. The 3×3 traditional convolution used in ResNet has inherent limitations: independent convolution kernels per channel lead to weak feature correlation; translation invariance simplifies computation but limits the receptive field; extracted features are single and fail to capture the differentiated behavior details

of elderly and children, seriously affecting detection accuracy in few-sample scenarios. Therefore, for the multi-target recognition needs of such spaces, the feature extraction mechanism of the backbone network must be reconstructed.

The optimized ResNet network is constructed based on the original ResNet structure with targeted modifications: all 3×3 traditional convolution modules are removed and uniformly replaced with deformable convolution, while the 1×1 convolution blocks at the head and tail of the main structure are retained. The retention of the 1×1 convolution blocks is to maintain channel mapping and feature fusion functions, ensuring dimensional matching of input and output features; replacing 3×3 convolution aims to break the limitations of traditional convolution and enhance feature extraction capabilities through the properties of deformable convolution, making it more adaptable to the recognition needs of multi-scale targets and complex behaviors in elderly-child shared spaces.

Specifically, the operational mechanism of deformable convolution is closely designed around the feature capture requirements of elderly-child shared spaces. For the input feature map $A \in R^{G \times Q \times Z}$, the feature vector $A_{uk} \in R^{Z}$ at a pixel (u,k) is first extracted, and then linearly transformed by the Ω function into a J^2-dimensional feature vector and unfolded into a J-dimensional kernel matrix G_{uk}. Assuming that the fully connected operation is denoted by $FC(\cdot)$, the ReLU activation function by $RELU(\cdot)$, and batch normalization by $BN(\cdot)$, then:

$$G_{uk} = \Omega(A) = FC\left(BN\left(RELU\left(FC(A)\right)\right)\right) \tag{1}$$

This process transforms the channel dimension information into spatial kernel information, allowing each pixel's convolution kernel to be associated with pixel features within the surrounding $[J/2]$ range. By performing point-wise multiplication, the channel information of the initial pixel is spread to the nearby spatial region, thereby collecting contextual information from a broader receptive field. This mechanism can effectively capture the spatial correlation of targets in elderly-child shared spaces, such as limb interaction features between elderly and children, compensating for the insufficient receptive field of traditional convolution. Let the output feature after deformation be B, then:

$$\left[a_1, a_2, ..., aj^2\right] \in \Re^{Z \times J^2} \otimes G_{uk} = \left[b_1, b_2, ..., bj^2\right] = S(a) \tag{2}$$

The final output is:

$$B = \sum_{u=1}^{j^2} b_u \tag{3}$$

The construction of the optimized ResNet network provides key support for multi-target behavior recognition in elderly-child shared spaces. Through channel dimension shared kernels, the deformable convolution performs multiplication and addition operations in a sliding window manner, summarizing context in a broad spatial structure, significantly expanding the receptive field and enhancing the feature capturing ability for elderly and child targets in complex scenes. At the same time, this design reduces parameter redundancy, improves feature correlation and richness, enables the learning of more robust behavior patterns from limited samples, effectively alleviates the overfitting problem in few-sample scenarios, and provides more precise high-level semantic features for subsequent region proposal and feature optimization, aiding efficient recognition of multi-target behaviors in elderly-child shared spaces. Figure 2 shows the schematic diagram of the deformable convolution in the optimized ResNet network.

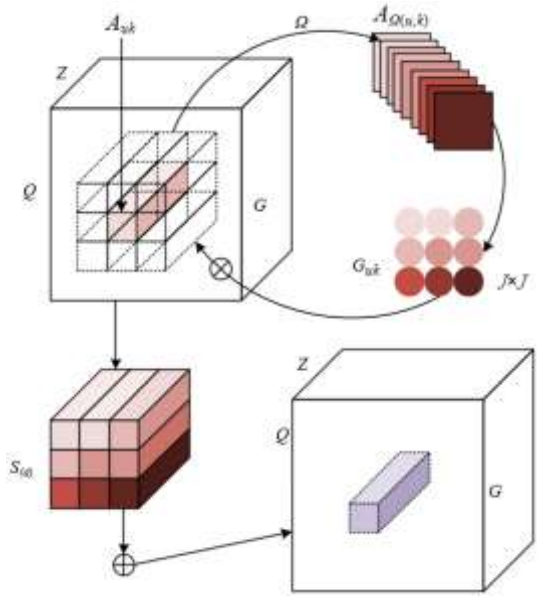

Figure 2. Schematic diagram of deformable convolution in the optimized ResNet network

2.2 Positive sample enhancement

The design of the positive sample enhancement branch module originates from the need to fully utilize object-level features in multi-target behavior recognition within elderly-child shared spaces. The network architecture is shown in Figure 3. In such spaces, the body size difference between elderly and children is significant, and frequent multi-target interactions easily lead to feature confusion. Meanwhile, the limited sample data contains insufficient positive samples at each scale, making it difficult for the region proposal network to accurately locate target regions. To solve this problem, the module enhances the positive sample features at each scale, strengthens the model's attention to elderly and child target regions, and compensates for the deficiencies of traditional region proposal networks in multi-scale target localization. This provides a more precise regional basis for subsequent behavior recognition and aligns with the research objective of improving feature enhancement effect and recognition accuracy.

The structural design of this module closely fits the multi-scale target characteristics of elderly-child shared spaces. First, the ground truth objects in the input image are cropped into multiple sizes from 32×32 to 800×800 to match the anchor sizes of each feature layer in the FPN, ensuring coverage of different body size scales of elderly and children in space and providing suitable inputs for positive sample feature enhancement at each scale. Second, after feature extraction via ResNet to different feature stages $p2$~$p6$, the corresponding scale feature maps are processed in two paths: one path is sent into the region proposal network, after 3×3 convolution and 1×1 convolution, used to calculate the foreground feature matrix of positive samples, which is aggregated and superimposed with the candidate feature matrix extracted by the optimized ResNet network to improve the object score of

elderly and child targets; the other path enters the spatial pyramid pooling network, is downsampled to 14×14, and fused with the filtered proposal regions, then input into the detection head for decoding, enhancing the feature information of positive samples at each scale. This dual-path design not only optimizes the quality of proposal boxes but also strengthens feature correlation, adapting to the complex distribution of multi-scale targets in elderly-child shared spaces.

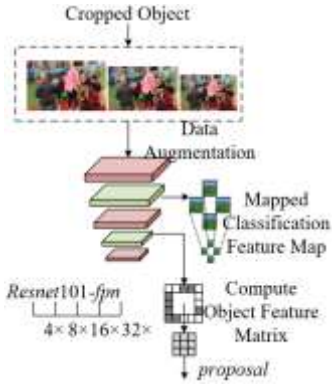

Figure 3. Network structure of the positive sample enhancement branch module

The positive sample enhancement branch module provides strong support for the model's multi-target behavior recognition through a targeted feature enhancement mechanism. By matching the cropped positive samples at different scales with anchors and integrating features, the module effectively supplements the scarce multi-scale positive sample features in elderly-child shared spaces, enabling the network to more accurately locate behavior regions of different scales such as child climbing or elderly standing up, and reducing localization deviation caused by target scale differences. At the same time, the aggregation and superposition of the foreground feature matrix and candidate feature matrix strengthen the feature saliency of elderly and child target regions, reduce the impact of background interference and target occlusion, provide more reliable feature input for subsequent behavior recognition modules, support the model to achieve more efficient multi-target behavior recognition in complex scenarios, and further promote the research objectives of feature enhancement and accurate recognition.

2.3 Loss function design

The construction of the supervised contrastive loss originates from addressing the problem of classification confusion of multi-target behaviors during few-shot fine-tuning in elderly-child shared spaces. In such spaces, some behavioral features of elderly and children targets exhibit similarity, and the sample data is limited, causing the model to easily confuse different categories of behaviors during fine-tuning. Traditional loss functions are difficult to accurately characterize instance-level differences in elderly and child behaviors, while supervised contrastive loss, inherited from contrastive learning ideas, can help the model learn more distinguishable high-level semantic features by strengthening intra-class feature aggregation and inter-class feature separation. This aligns with the research goal of improving multi-target behavior recognition accuracy under few-shot conditions. Its specific construction method is closely centered around behavioral features in elderly-child shared spaces. In feature processing, the proposed box features extracted by the spatial pyramid pooling network are flattened and mapped through a fully connected layer to a 1024-dimensional feature vector p, and then $L2$-normalized and encoded to reduce the dimension to a 128-dimensional embedding feature $p^{\sim}$, which simplifies the comparison complexity of elderly and child behavior features and adapts to the feature dimension requirements of multi-target behavior in this space. In similarity measurement, the cosine projection space is used to compute the similarity between embedding features to characterize the class belonging probability of elderly and child behaviors. That is, for significantly different behaviors such as child climbing and elderly falling, the similarity of different-class features is reduced; for same-class behaviors, the similarity is increased, clarifying intra-class compactness and inter-class separability. In loss function design, when the total similarity of same-class features is larger, the loss becomes smaller; when the total similarity of different-class features is smaller, the loss also becomes smaller, guiding the model to focus on behavioral differences of multi-targets in elderly-child shared spaces. Figure 4 shows a schematic diagram of the loss function design principle. Assume that any $o^{\sim}_u$ and its same-class features are represented by $o^{\sim}_k$, and different-class features are represented by $o^{\sim}_j$; the total number of embedding features $o^{\sim}$ is denoted as V, the ground truth corresponding to feature $o^{\sim}_u$ is denoted as b_u, the number of features of b_u is V_{bu}, and δ is a hyperparameter greater than P. Based on the above, the loss functions are designed as:

$$LOSS_{EX} = \frac{1}{V}\sum_{u=1}^{V} LOSS_{ou} \tag{4}$$

$$LOSS_{ou} = \frac{-1}{V_{bu}-1}\sum_{k=1,k\neq u}^{V} II\{b_u = b_k\}\cdot \log \frac{\exp\left(\tilde{o}_u \cdot \frac{\tilde{o}_k}{\delta}\right)}{\sum_{j=1}^{V} II_{j\neq u} \cdot \exp\left(\tilde{o}_u \cdot \frac{\tilde{o}_j}{\delta}\right)} \tag{5}$$

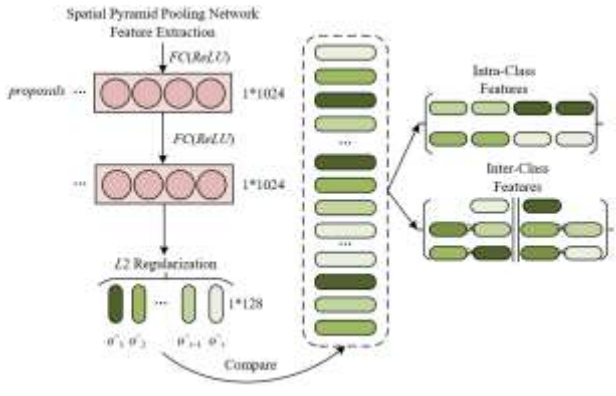

Figure 4. Schematic diagram of the loss function design principle

In summary, the total loss function of the model is defined as consisting of the binary cross-entropy loss $LOSS_{EOV}$ for optimizing foreground objects, the binary cross-entropy loss $LOSS_{zmt}$ for foreground object classification, the smooth L1

loss $LOSS_{REG}$, and $LOSS_{EX}$:

$$LOSS_{\text{total}} = LOSS_{EOV} + LOSS_{zmt} + LOSS_{REG} + \frac{1}{2} LOSS_{EX} \tag{6}$$

The supervised contrastive loss plays a key role in the model. It, together with the binary cross-entropy loss and smooth L1 loss, constitutes the total loss function to jointly optimize the model performance. In few-shot scenarios, it can effectively reduce the classification confusion of elderly and child behaviors, such as avoiding misjudging elderly people sitting down slowly as children squatting, and enhance the model's ability to distinguish similar behaviors. Meanwhile, by prompting the model to focus on features of the same-class behavior, it strengthens the semantic discriminability of multi-target behavior features in elderly-child shared spaces, provides loss constraints for the model to learn robust behavior patterns from limited samples, and helps to achieve the research goals of feature enhancement and precise recognition, thereby improving the overall reliability of multi-target behavior recognition.

2.4 Model training

This paper introduces a balanced fine-tuning method to solve the problem of sample imbalance in deep network training under few-shot conditions in elderly-child shared spaces, aiming to improve the accuracy of multi-target behavior recognition under few-shot scenarios. In such spaces, behavior samples specific to elderly and children are often scarce, while common behavior samples are relatively sufficient. During fine-tuning, new-class positive samples are likely to be erroneously suppressed by non-maximum suppression due to low foreground confidence in the region proposal network. At the same time, excessive background negative samples interfere with the model's learning of effective features. Traditional fine-tuning strategies cannot balance the training weights of base classes and new classes, making it difficult for the network to accurately capture specific behavioral features of the elderly and children. Therefore, a targeted balanced mechanism is needed to optimize the training process.

To address the problems of positive sample filtering and negative sample interference, the balanced fine-tuning method achieves optimization through joint training and sample selection strategies. The region proposal network and the spatial pyramid pooling network feature extractor are jointly trained under elderly-child shared space object supervision, which can double the number of candidate boxes that pass non-maximum suppression, ensuring that rare new-class positive samples such as elderly falling and child climbing are retained. Meanwhile, the number of candidate boxes used for loss calculation in the spatial pyramid pooling network is halved to reduce interference from background-only negative samples, enabling the model to focus more on effective behavior features of elderly and child targets. Assume that the original gradients obtained from training the new class and base class are h_v and h_y, and the updated gradient values are $\tilde{h}$, $\tilde{h}_y$. The following formulas give the optimization objective function of fine-tuning training:

$$MIN\ S\left(\tilde{h}_y, \tilde{h}_v\right) = \frac{1}{2}\left\|h_v - \tilde{h}_v\right\|_2^2 + \frac{1}{2}\left\|h_y - \tilde{h}_y\right\|_2^2 \tag{7}$$

$$\tilde{h}_v = h_v - \left(\frac{h_v^S \cdot h_y}{h_y^S \cdot h_y}\right) \cdot h_y, \tilde{h}_y = h_y - \left(\frac{h_y^S \cdot h_v}{h_v^S \cdot h_v}\right) \cdot h_v \tag{8}$$

In elderly-child shared spaces, the base class data is large while the new-class samples are scarce. Traditional gradient updates tend to favor base classes, making it difficult for the network to remember new-class features of specific elderly and child behaviors. Balanced gradients can adaptively reweight the gradients of the two types of data, allowing the network to quickly acquire new-class information from limited samples and efficiently generalize to various behaviors of elderly and child groups. Ultimately, this improves the stability and accuracy of the model's multi-target behavior recognition under few-shot conditions. Specifically, this paper adopts balanced gradient descent instead of traditional stochastic gradient descent, computing the weighted average of base class gradient h_y and new class gradient h_v as the balanced gradient $\tilde{h}$, with the calculation formula as follows:

$$\tilde{h} = \frac{1}{2}\left(1 - \frac{h_v^S \cdot h_y}{h_y^S \cdot h_y}\right) \cdot h_y + \frac{1}{2}\left(1 - \frac{h_y^S \cdot h_v}{h_v^S \cdot h_v}\right) \cdot h_v \tag{9}$$

3. ELDERLY-CHILD SHARED SPACE SAFETY MONITORING STRATEGY BASED ON MULTI-TARGET BEHAVIOR RECOGNITION IN VIDEO IMAGES

After completing multi-target behavior recognition in video images, the primary strategy is to construct a behavior-feature-based risk assessment and graded response mechanism. In view of the particularity of elderly-child shared spaces, recognized behaviors such as elderly falling, children climbing dangerous objects, and elderly-children separation are matched with a predefined safety rule base. By combining behavior duration, target location, and environmental risk coefficient, the behavior risk value is quantified to achieve risk level classification. For high-risk behaviors such as elderly falling with head impact, the system immediately triggers a first-level response, automatically pushing alerts with real-time images and precise location to the management terminal and the family members' mobile phones, and linking on-site sound and light alarm devices. For medium-risk behaviors such as children approaching the balcony edge, a second-level response is initiated, where the monitoring center sends a prompt message to on-site staff and activates the camera to track the target's movement. For low-risk behaviors such as elderly sitting still for a long time, the behavior data is recorded for trend analysis to provide a reference for subsequent care. This strategy achieves a fast closed-loop process from recognition to response by accurately matching the behavioral risk characteristics of elderly and child groups.

Secondly, a multidimensional real-time linkage and intelligent intervention system is established to strengthen the initiative and effectiveness of safety monitoring. Based on multi-target behavior recognition results, the system links physical facilities and management platforms within the space in real time, such as intelligent access control, emergency lighting, and automatic handrails. When a child is identified walking alone towards the exit, the access control is automatically delayed in closing and a warning is pushed; if an elderly person is detected walking in a slippery area, the

ground warning light is activated and a voice reminder is broadcast. At the same time, a behavior habit model for elderly and children is constructed based on historical behavior data. Abnormal behaviors such as a significant reduction in the elderly's daily activity time are pre-warned in advance to assist management personnel in predicting potential risks. In addition, the system supports remote intervention functions. Management personnel can use the monitoring platform to conduct voice guidance for abnormal behavior areas, forming a full-process safety management loop of "recognition-analysis-intervention-feedback", effectively improving the safety assurance capability of elderly-child shared spaces.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. Experimental data comparison on training set / %

Model/Method	Average Detection Accuracy for New Classes			
	1 Target	2 Targets	3 Targets	4 Targets
Mask R-CNN	4.3	22.3	31.5	41.5
Model-Agnostic Meta-Learning	17.8	31.5	52.4	58.6
SENet	15.6	28.6	48.6	62.3
NAS-FPN	18.5	31.4	51.2	56.7
Feature Reconstruction Detector	17.5	32.5	51.4	62.5
Ours	22.3	34.8	52.9	62.8

From the experimental data comparison in Table 1 and Table 2, it can be seen that on both the training set and the test set, the method proposed in this paper achieves significantly better average detection accuracy for new classes under single-target and multi-target scenarios compared with the baseline models including Mask R-CNN, Model-Agnostic Meta-Learning, SENet, NAS-FPN, and Feature Reconstruction Detector. Specifically, in the training set, the accuracy of Our Method reaches 62.8% in the 4-target scenario, far exceeding the 41.5% of Mask R-CNN. In the test set, ours achieves an accuracy of 52.3% in the 4-target scenario, also outperforming other models. This indicates that by optimizing the feature extraction network, the proposed method enhances the discriminability of elderly and child behavior features, making feature extraction and recognition more robust in complex multi-target scenarios, effectively improving the accuracy of multi-target behavior detection. The multi-target accuracy in both the training and test sets steadily increases with the number of targets and always remains higher than that of the comparison models. In summary, the experimental data fully verify the effectiveness of the proposed method in multi-target behavior recognition: it performs excellently not only in single-target scenarios but also shows stronger detection capability in complex multi-target elderly-child shared scenes, significantly surpassing existing comparison methods and providing high-precision and robust technical support for safety monitoring in elderly-child shared spaces.

Table 2. Experimental data comparison on test set / %

Model/Method	Average Detection Accuracy for New Classes			
	1 Target	2 Targets	3 Targets	4 Targets
Mask R-CNN	8.3	15.6	24.5	27.8
Model-Agnostic Meta-Learning	15.9	31.4	32.8	37.5
SENet	23.4	36.8	43.2	52.6
NAS-FPN	15.8	31.5	35.6	41.9
Feature Reconstruction Detector	22.3	31.9	36.8	47.5
Ours	26.9	38.6	44.6	52.3

Table 3. Ablation experiments of each module

No.	Optimized ResNet Network	Positive Sample Enhancement Branch	Supervised Contrastive Loss	Average Detection Accuracy of New Classes			
				1 Target	2 Targets	3 Targets	4 Targets
1	×	×	×	8.3	15.9	24.5	27.5
2	√	×	×	9.5	16.8	25.3	31.6
3	√	√	×	24.6	36.2	41.8	52.6
4	√	√	√	26.8	38.7	44.5	52.7

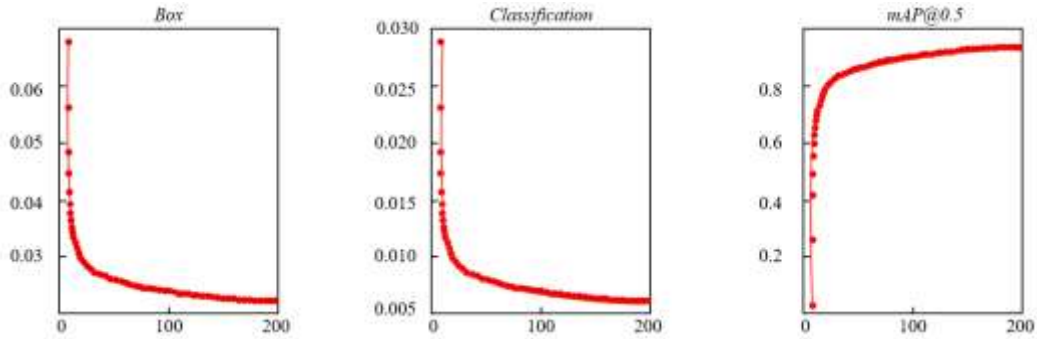

Figure 5. Trends of loss and mAP @ 0.5 of the proposed method

Through the ablation experiment analysis in Table 3, the proposed method demonstrates significant effectiveness under the collaborative optimization of multiple modules. In the experiment, No. 1 represents the baseline without any optimization modules, and its average detection accuracy for new classes is low under both single and multi-target scenarios, reflecting the insufficient capability of the baseline model to extract and distinguish behavior features of elderly

and children. No. 2 only applies the optimized ResNet network, with a slight improvement in accuracy, verifying the fundamental role of the optimized feature extraction network in behavior recognition. After adding the positive sample enhancement branch in No. 3, the accuracy increases significantly, indicating that the enhancement of positive samples effectively strengthens the discriminability of elderly and child behavior features and improves the robustness of recognition in complex multi-target scenarios. No. 4 further introduces supervised contrastive loss, with accuracy continuing to improve, reflecting the continuous optimization of model performance through module collaboration. The experimental results show that the feature-enhanced multi-target behavior recognition framework constructed through the optimized ResNet network, positive sample enhancement branch, and supervised contrastive loss, demonstrates layered progression and synergistic improvement across modules, significantly enhancing detection accuracy in both single and multi-target scenarios. The ablation experiments fully validate the effectiveness of the proposed method: the multi-module optimization based on feature enhancement greatly improves the accuracy and robustness of multi-target behavior recognition.

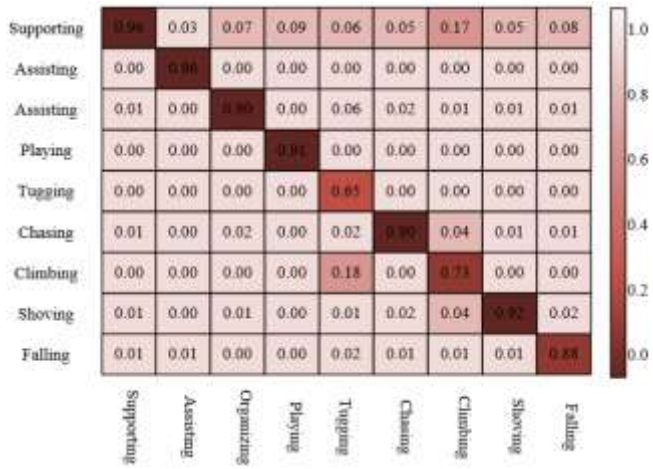

Figure 6. Confusion matrix

Through the trend analysis of Loss and mAP in Figure 5, the proposed method shows significant performance optimization and convergence characteristics during training, fully validating its effectiveness. In the left Box subplot, the bounding box regression loss rapidly decreases with iterations and stabilizes, indicating that the model's ability to predict the positions of elderly and child targets is continuously improved under the effect of feature enhancement technology, accurately locating multi-target behavior subjects in complex scenes, and providing reliable spatial location information for subsequent safety monitoring. In the middle Classification subplot, the rapid decline of classification loss reflects the model's significantly enhanced ability to identify elderly and child behavior categories, thanks to the strengthening of feature discriminability, enabling the model to accurately classify behavior types in elderly-child scenarios with coexisting multiple targets, providing key semantic information for safety risk assessment. In the right mAP@0.5 subplot, the mean average precision quickly rises and reaches a high value close to 0.9, reflecting the effect of collaborative optimization between Box and Classification loss, proving that the overall performance of the proposed method in multi-target behavior recognition far exceeds that of the baseline model.

Through the analysis of the confusion matrix in Figure 6, the proposed method shows excellent performance in multi-target behavior recognition in elderly-child shared spaces. The diagonal elements of the matrix are all at high levels, such as "supporting" 0.96, "assisting" 0.96, "playing" 0.91, "falling" 0.88, indicating significant correct recognition rates for core behaviors including safe interaction, playing, and high-risk abnormal behaviors, and reflecting the effectiveness of feature enhancement technology in distinguishing elderly and child behavior features. For example, the high recognition rate of positive safety behaviors such as “supporting” and “assisting” validates the model’s accurate capture of elderly-child mutual aid scenarios; the accurate classification of high-risk behaviors such as “falling” and “pushing” ensures timely warning of abnormal behaviors, providing a reliable behavior recognition basis for safety monitoring strategies. Among the off-diagonal elements, although there are some misclassifications, the overall error is controllable, and the recognition accuracy of key safety-related behaviors still meets practical application requirements. For example, “falling” as an emergency risk behavior, with an accuracy rate of 0.88, can effectively trigger emergency responses. Combined with the research content, the feature enhancement module improves the robustness of the model in complex scenarios by strengthening the discriminability of behavior features, making the confusion matrix show a desirable distribution of "high diagonal, low off-diagonal errors", providing high-precision behavior classification results for the safety risk assessment model. In summary, the confusion matrix visually reflects the high recognition accuracy of the proposed method for multi-target behaviors in elderly-child shared scenarios, especially for precise classification of safety-related behaviors, fully proving the collaborative effectiveness of the technical approach and effectively improving the safety monitoring capability of elderly-child shared spaces.

Through the analysis of the PR curve in Figure 7, the proposed method shows outstanding performance in multi-target behavior recognition in elderly-child shared spaces. The PR curves of various behavior categories are close to the upper left corner, indicating that precision remains high under different recall rates. The average precision (AP) of safe interaction behaviors such as "playing", "assisting", "organizing", as well as high-risk behaviors such as "pushing" and "falling", all exceed 0.9, reflecting the precise extraction and discrimination capability of the feature enhancement technology for elderly and child behavior features. The overall mAP@0.5 of all categories reaches 0.911, further verifying the model’s comprehensive recognition performance under multi-target scenarios. By optimizing the feature extraction network, the model is able to classify safe interaction and risky behaviors with high accuracy in complex elderly-child coexisting scenarios, providing a reliable behavior recognition basis for safety monitoring strategies. Based on the high-accuracy behavior recognition results, the safety risk assessment model can capture abnormal behaviors in real time and trigger emergency responses, achieving proactive safety prevention and control. The excellent performance of the PR curve fully proves the effectiveness of the proposed method: from feature enhancement to behavior recognition, the technical optimization enables the model to achieve both accuracy and robustness in multi-target behavior detection in elderly-child shared spaces, strongly supporting the

implementation of safety monitoring strategies and providing a practical solution for the safety protection of elderly and child groups.

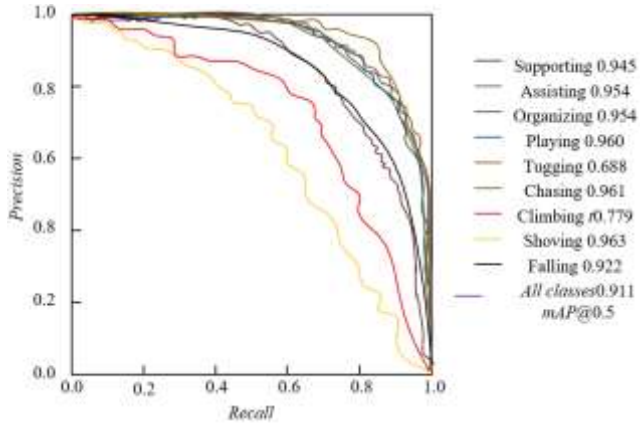

Figure 7. PR curve

5. CONCLUSION

This paper focused on safety monitoring in elderly-child shared spaces and proposed a technical solution based on video image multi-target behavior recognition. In terms of core technologies, by optimizing the feature extraction network, positive sample enhancement branch, and supervised contrastive loss, the discriminability of elderly and child behavior features was strengthened, significantly improving the accuracy and robustness of multi-target behavior recognition. Experimental data show that the model achieved high recognition accuracy for core behaviors such as "supporting," "assisting," and "falling," effectively solving the problem of behavior classification in complex scenarios. At the application level, based on the high-accuracy behavior recognition results, a safety risk assessment model was constructed, realizing real-time warning and emergency response linkage for abnormal behaviors, covering scenarios such as elderly-child interaction and dangerous coordination, providing practical technical support for actual safety monitoring. The research results not only verified the effectiveness of the technical chain of "feature enhancement–behavior recognition–safety monitoring," but also provided an innovative solution for intelligent safety management in elderly-child shared spaces, with important application value, promoting the implementation of intelligent monitoring technology in the field of vulnerable group protection.

Although the proposed method performs well, there are still limitations: First, the behavior recognition accuracy under extreme environments needs to be optimized, and the misclassification rate of some low-frequency behaviors is relatively high, indicating that the model's generalization ability in complex dynamic scenes needs to be enhanced; second, the model has high computational complexity, posing challenges for real-time deployment. Future research can break through from the following directions: (1) Multimodal fusion: combining audio and sensor data to construct a multimodal feature space, improving recognition robustness in complex scenarios and solving the limitations of single visual information; (2) Model lightweighting and real-time implementation: exploring lightweight networks, combining model compression and edge computing to achieve low-latency, high-cost-performance real-time monitoring deployment, adapting to the hardware requirements of actual scenarios; (3) Dataset and scene optimization: constructing a richer elderly-child behavior dataset, conducting customized training for specific scenarios such as nursing homes and childcare centers, enhancing the method's scenario adaptability, and promoting the deepening of technology from "general-purpose" to "scenario-specific." Through research in the above directions, the technical system can be further improved, the practicality and universality of the method can be enhanced, helping elderly-child safety monitoring technology move from laboratory verification to large-scale application, contributing more valuable solutions to the field of intelligent safety monitoring, and ultimately achieving active, intelligent, and precise safety protection in elderly-child shared spaces.

ACKNOWLEDGEMENTS

This paper was supported by Henan Province's Science and Technology Development Plan for 2025 under the project "Research on the Construction of an Integrated Service System for the Elderly and Children in Henan Province" (Grant No.: 252400410020).

REFERENCES

[1] Ristl, C., Korlat, S., Rupprecht, F.S., Burgstaller, A., Nikitin, J. (2025). Self-perceptions of aging and social goals. Psychology and Aging, 40(4): 413-420. https://doi.org/10.1037/pag0000881

[2] González, N.T., Machanda, Z., Thompson, M.E. (2023). Age-related social selectivity: An adaptive lens on a later life social phenotype. Neuroscience & Biobehavioral Reviews, 152: 105294. https://doi.org/10.1016/j.neubiorev.2023.105294

[3] Chapin, R., Nelson-Becker, H., Gordon, T., Landry, S.T., Chapin Jr, W.B. (2007). Responding to the Hartford geriatric social work initiative: A multilevel community approach to building aging competency. Journal of Gerontological Social Work, 50(1-2): 59-74. https://doi.org/10.1300/J083v50n01_05

[4] Jackson, K.E., Krishnaswami, S., McPheeters, M. (2011). Unmet health care needs in children with cerebral palsy: A cross-sectional study. Research in Developmental Disabilities, 32(6): 2714-2723. https://doi.org/10.1016/j.ridd.2011.05.040

[5] Thsitake, R.S., Pengpid, S., Peltzer, K. (2013). Knowledge and experiences of child care workers regarding care and management of children with special needs in Gauteng, South Africa. Journal of Child & Adolescent Mental Health, 25(2): 131-138. https://doi.org/10.2989/17280583.2013.801845

[6] Sundevall, E.P., Jansson, M. (2020). Inclusive parks across ages: Multifunction and urban open space management for children, adolescents, and the elderly. International Journal of Environmental Research and Public Health, 17(24): 9357. https://doi.org/10.3390/ijerph17249357

[7] Boaz, R.F., Hu, J., Ye, Y. (1999). The transfer of resources from middle-aged children to functionally limited elderly parents: Providing time, giving money, sharing space. The Gerontologist, 39(6): 648-657. https://doi.org/10.1093/geront/39.6.648

[8] Zheng, L., Yang, C., Zhang, W., Cai, X., et al. (2013). Comparison of multi-space infections of the head and neck in the elderly and non-elderly: Part I the descriptive data. Journal of Cranio-Maxillofacial Surgery, 41(8): e208-e212. https://doi.org/10.1016/j.jcms.2013.01.020

[9] Aldabbagh, L.A. (2024). Implementing parks in Mosul City-(residential neighbourhood level). Journal of Urban Development and Management, 3(4): 227-240. https://doi.org/10.56578/judm030402

[10] Utari, D.T., Hendradewa, A.P., Bella, M.A. (2025). Optimized YOLO approach for drowsiness detection in automotive safety: Parameter tuning and facial expression analysis. International Journal of Transport Development and Integration, 9(1): 189-196. https://doi.org/10.18280/ijtdi.090118

[11] Martínez-Ballesté, A., Rashwan, H., Puig, D., Solanas, A. (2018). Design and implementation of a secure and trustworthy platform for privacy-aware video surveillance. International Journal of Information Security, 17(3): 279-290. https://doi.org/10.1007/s10207-017-0370-4

[12] Nguyen, C., Feng, W.C., Liu, F. (2016). Hotspot: Making computer vision more effective for human video surveillance. Information Visualization, 15(4): 273-285. https://doi.org/10.1177/1473871616630015

[13] Vatambeti, R., Damera, V.K. (2022). Gait based person identification using deep learning model of generative adversarial network. Acadlore Transactions on AI and Machine Learning, 1(2): 90-100. https://doi.org/10.56578/ataiml010203

[14] Fitzpatrick, G., Lipovetzky, N., Papasimeon, M., Ramirez, M., Vered, M. (2021). Behaviour recognition with kinodynamic planning over continuous domains. Frontiers in Artificial Intelligence, 4: 717003. https://doi.org/10.3389/frai.2021.717003

[15] Lee, Y.C., Lee, S.Y., Kim, B., Kim, D.Y. (2024). GLBRF: Group-based lightweight human behavior recognition framework in video camera. Applied Sciences, 14(6): 2424. https://doi.org/10.3390/app14062424

[16] Yoon, K., Song, Y.M., Jeon, M. (2018). Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. IET Image Processing, 12(7): 1175-1184. https://doi.org/10.1049/iet-ipr.2017.1244

[17] Liang, M., Kim, D.Y., Kai, X. (2015). Multi-Bernoulli filter for target tracking with multi-static Doppler only measurement. Signal Processing, 108: 102-110. https://doi.org/10.1016/j.sigpro.2014.09.013

[18] Ko, K.E., Sim, K.B. (2018). Deep convolutional framework for abnormal behavior detection in a smart surveillance system. Engineering Applications of Artificial Intelligence, 67: 226-234. https://doi.org/10.1016/j.engappai.2017.10.001

[19] Li, J., Huang, Q., Du, Y., Zhen, X., Chen, S., Shao, L. (2021). Variational abnormal behavior detection with motion consistency. IEEE Transactions on Image Processing, 31: 275-286. https://doi.org/10.1109/TIP.2021.3130545

[20] Feizi, A. (2020). Hierarchical detection of abnormal behaviors in video surveillance through modeling normal behaviors based on AUC maximization. Soft Computing-A Fusion of Foundations, Methodologies & Applications, 24(14): 10401-10413. https://doi.org/10.1007/s00500-019-04544-9