Enhanced 3D Residual Network for Human Fall Detection in Video Surveillance

Suyuan Li[1], Xin Song[1,2,*], Jing Cao[1,2,3] and Siyang Xu[1,2]
[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
[2] Engineering Optimization and Smart Antenna Institute, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China
[3] School of Mathematics and Information Science and Technology, Hebei Normal University of Science and Technology, Qinhuangdao 066004, China
[e-mail: 2010649@stu.neu.edu.cn, sxin78916@neuq.edu.cn, owenjing@sina.com, 1810584@stu.neu.edu.cn]
[*]Corresponding author: Xin Song

Received March 16, 2022; revised September 15, 2022; accepted November 22, 2022; published December 31, 2022

Abstract

In the public healthcare, a computational system that can automatically and efficiently detect and classify falls from a video sequence has significant potential. With the advancement of deep learning, which can extract temporal and spatial information, has become more widespread. However, traditional 3D CNNs that usually adopt shallow networks cannot obtain higher recognition accuracy than deeper networks. Additionally, some experiences of neural network show that the problem of gradient explosions occurs with increasing the network layers. As a result, an enhanced three-dimensional ResNet-based method for fall detection (3D-ERes-FD) is proposed to directly extract spatio-temporal features to address these issues. In our method, a 50-layer 3D residual network is used to deepen the network for improving fall recognition accuracy. Furthermore, enhanced residual units with four convolutional layers are developed to efficiently reduce the number of parameters and increase the depth of the network. According to the experimental results, the proposed method outperformed several state-of-the-art methods.

Keywords: Video surveillance, fall detection, deep learning, residual network, 3D CNN.

1. Introduction

Fall is an involuntary, unintended significant postural change that occurs in the elderly, patients, athletes, laborers, and even healthy persons, resulting in deterioration of human health. [1]. Due to the obvious physical weakness that occurs with aging, the elderly account for the majority of fatal falls. According to a report, about 1.6 million older individuals in the United States suffer from injuries of falls each year [2]. Nowadays, fall is the second common cause of mortality. Fall will not only cause physical and psychological harm to the elderly, but also has a serious impact on the family and society, as shown in **Fig. 1**. As a result, developing a real-time and effective system to detect fall event is very critical for the elderly. Recently, numerous researches have been conducted in order to build an intelligent monitoring system for the elderly that can detect falls automatically and instantaneously. Nowadays, fall detection methods based on vision have gained more attention than those based on wearable sensors and ambient sensors, due to greater need for user-friendliness, such as simplicity of use, non-invasiveness, and minimally affecting the user's regular activities. As a consequence, fall detection systems based on vision will be more universality and feasibility in the future, due to comprehensive monitoring information, non-contact surveillance, and a monitoring environment with no electromagnetic interference [3].

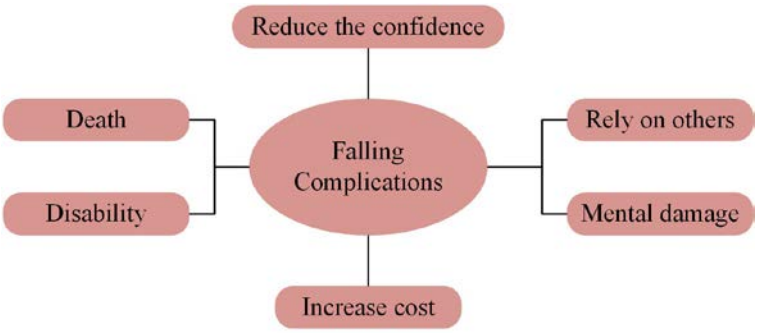


Fig. 1. Falling Complications

Traditional methods based on hand-crafted features and novel methods based on deep learning are two kinds of vision-based fall detection systems. Compared with traditional methods, CNNs [4] can extract features automatically and obtain greater recognition accuracy. Thus, deeper networks have increasingly been applied to recognize a fall event from a video. However, traditional 3D CNN that usually adopts shallow networks can't obtain higher recognition accuracy than deeper networks. Simultaneously, as the depth of 3D CNN rises, more model parameters and gradient explosions are included, which will increase the complexity of the constructed model.

In the paper, we propose a novel three-dimensional ResNet-based method for fall detection (3D-ERes-FD) to address the aforementioned issues. To efficiently retain temporal information in a fall video and improve detection performance, 3D ResNet is utilized to extract spatio-temporal features and further increases the feature's representativeness in the network. Besides, enhanced residual unites are constructed to minimize the number of parameters as the network layer increases.

The remaining paper is organized as follows. Section 2 briefly reviews the related work. Section 3 outlines the details of the proposed fall detection method using the enhanced residual

units. In Section 4, different experimental results compared with other fall detection methods are provided. Finally, Section 5 concludes the paper.

2. Related Work

Fall detection is a critical and attractive research topic in the field of public healthcare to improve healthcare and medical services. Generally, fall detection methods are divided into three categories based on the equipment involved: wearable sensors, ambient sensors, and cameras.

Wearable sensors, such as accelerometers, gyroscopes, and electromyography, are generally connected to the chest, waist, or wrist to gather data. Recently, to decrease false alarms, a comprehensive fall detection system based on accelerometers and smart phones is suggested, the threshold-based technique and multiple kernel learning support vector machine are used [5]. To develop a wearable airbag, accelerometer and gyroscope sensor are utilized to collect both acceleration and angular velocity signals to trigger inflation of the airbag [6]. Additionally, the feature extraction and pattern recognition of surface electromyography (sEMG) can be adopted to detect muscles electrical changes in falls [7]. Although the wearable sensors are low cost and convenient, the permanent wearable manner is not pleasant for people.

Ambient sensors primarily gather data from monitoring surroundings, including sound, vibration, pressure and light intensity. In [8], a novel floor acoustic sensor is utilized to record acoustic waves transmitted through the floors, and then the acoustic waves can be classified by a two class Support Vector Machine for detecting a fall event. Due to the fact that pain can be immediately reflected through sound, the hidden Markov model based on component analysis (HMM-CA) is adopted to separate overlapping sound signals [9]. Moreover, because the pressure sensor has reliable operation and stable performance, a novel smart sensing technique with piezoresistive pressure sensors is developed to trigger an alarm when a fall event occurs [10]. Despite the ambient sensors can protect individual privacy, all surfaces need to be generally covered with these sensors, which is laborious and easily exposed to noise.

Since video cameras have numerous advantages over wearable sensors and ambient sensors, including rich information, non-invasive collection, no electromagnetic interference, pleasant user experience, low cost, and so on, fall detection methods based on cameras are widely adopted in recent years. Generally, cameras are used to separate human subjects from scenes collected by RGB cameras, Kinect cameras, thermal sensors, or even numerous cameras [11]. Traditionally, fall detection methods based on hand-crafted features mainly focus on tracing the head trajectories, body shape, or body posture for detecting a fall event. These methods about tracing the head trajectories are premised on the theory that vertical motion is stronger than horizontal motion when fall occurs. Moreover, the tracked major joints of the human body are analyzed by a pose-invariant randomized decision tree, then the 3D trajectory of the head joint is input into the SVM classifier to determine falls [12]. Due to the significance of the motion information, motion history image and code-book background subtraction are applied to detect large movement, and particle filters based on the magnitude of the movement information can track the head. Then, falls can be detected through the three dimensional horizontal and vertical velocities of the head [13]. In contrast to tracing head trajectories, methods based on human body shape and posture concentrate on the transition from standing to laying when a fall event happens. Hidden Markov Model is applied to retrieve numerous characteristics from the silhouette, including the height of the bounding box and the magnitude of the motion vector, to effectively identify a fall event [14]. Furthermore, several cameras are used to extract the 3D shape of the person, and the multi-camera vision system for recognizing

and tracking humans uses a silhouette warping technique to convey visual information between overlapping cameras [15]. However, all known video-based methods involving extracting the subject first are prone to being affected by background noise.

Recently, fall detection systems are based on machine learning algorithms and deep learning algorithms, such as SVM and CNN. In [16], 2D CNN is straightway utilized to obtain human shape features from the frames for distinguishing a fall event. Whereas, the information between video frames cannot be fully utilized by CNN, which will have an impact on detection performance. Thus, the fall detection methods based on recurrent neural network (RNN) are very effective for sequential data to get more attention gradually. In [17], RNN and Long Short Term Memory (LSTM) are employed to process skeleton extracted by CNN, and the output of RNN can be utilized to distinguish a fall event. Besides, optical flow technique can also be utilized to better portray the relationship between video frames. To utilize preprocessed video frames, optical flow images can be input into wide residual network to extract features for detecting a fall event [18]. Optical flow images are used as input to CNN, along with the next three-step training process, to simulate video motion and make the system scenario independent [19]. Except for optical flow, many fall detection methods attempt to combine different preprocessing method to improve detection performance. In [20], a multi-stream CNN architecture for human-related areas is presented, which can encode appearance, motion, and the captured tubes of the human-related regions. However, the above methods mainly adopt 2D CNN as extractor, which cannot make full use of temporal correlation. As a consequence, 3D network is developed to effectively capture temporal and spatial features over each frame. In [21], 3D CNN is first adopted to capture the motion information contained in consecutive frames, which can extract features from both the spatial and temporal dimensions. Simultaneously, with the development of LSTM, 3D CNN is utilized to obtain motion features from temporal sequence, and then a spatial visual attention strategy based on LSTM is adopted to detect falls [22]. To enrich more input, multi-stream visual characteristics are integrated into 3D CNN network for action identification in clipped videos [23]. Nevertheless, traditional 3D CNN that usually adopts shallow networks cannot obtain higher recognition accuracy than deeper networks. Furthermore, the problem of gradient explosion will occur with blindly increasing the depth of the network.

To overcome the above limitation, a fall detection method based on enhanced 3D ResNet50 is presented to directly extract spatio-temporal information from videos for detecting falls. In our method, a 50-layer 3D residual network is adopted to deepen network for obtaining a better fall recognition accuracy. Additionally, enhanced residual units with four convolutional layers are offered to reduce the number of parameters while increasing the network's layer.

3. Proposed Method

In this paper, a novel method is proposed to detect fall events in video effectively, which is based on residual neural network using spatio-temporal 3D kernels. The architecture of the proposed 3D-ERes-FD method is shown in **Fig. 2**. Firstly, original video is processed to obtain cropped sequence, which can eliminate some redundant information. Then, 3D residual CNN is considered as extractor to obtain features from cropped sequence. In the end, the softmax classifier is adopted to classify the optimal features for distinguishing fall events.

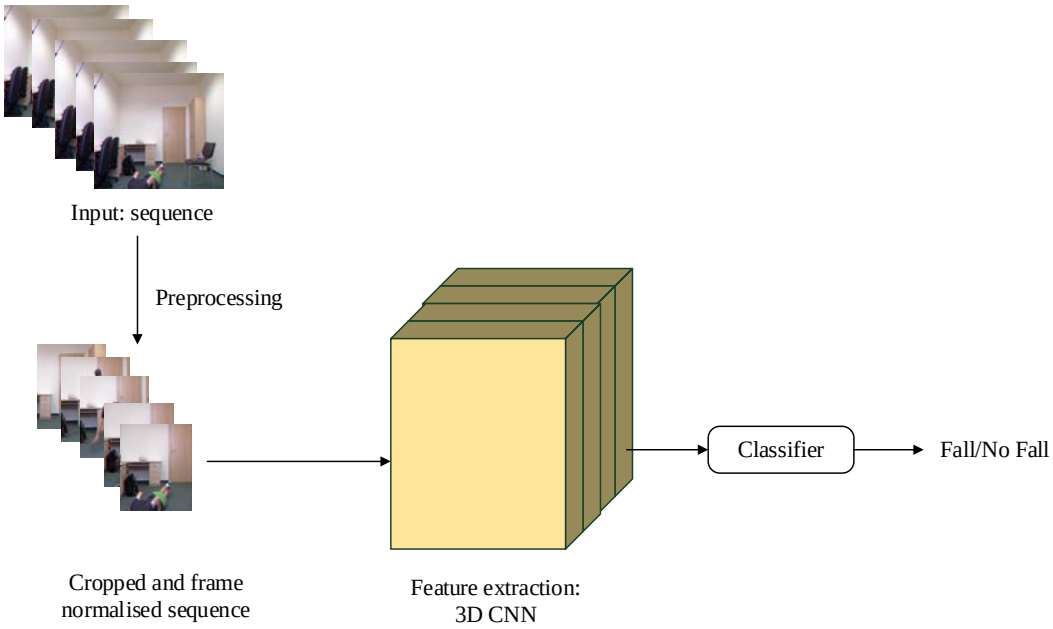


Fig. 2. Architecture of the proposed 3D-ERes-FD

3.1 3D Convolutional Neural Network

In 2D convolutional neural network, 2D convolution is performed to extract features from local neighborhood on feature maps in the previous layer. As shown in **Fig. 3**, the convolution kernel (filter) slides over the spatial dimensions of the input image, the slide window is convolved with values in the convolution kernel each time to obtain a value of output. Formally, 2D CNN is as follows:

$$v_{ij}^{xy} = f\left(\sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} + b_{ij}\right) \tag{1}$$

where the value at location (x, y) of the j_{th} feature map in the i_{th} layer can be represented by v_{ij}^{xy}, the activation function can be represented by $f(\bullet)$, the j_{th} bias in the i_{th} layer can be represented by b_{ij}, the vertical (spatial) and horizontal (spatial) extents can be represented P_i, Q_i, the index of the set of feature maps from the $(i-1)_{th}$ layer is presented by m, the value of the filter cube at location (p, q) connected to the m_{th} feature map in the previous layer is represented by w_{ijm}^{pq}, and the value at location $(x+p, y+q)$ in the m_{th} feature map in the $(i-1)_{th}$ layer is represented by $v_{(i-1)m}^{(x+p)(y+q)}$.

In 2D CNN, convolutions are implemented to the 2D feature maps to obtain features based solely on the spatial dimension. In video analysis problems, it's not appropriate to collect motion information stored in numerous adjacent frames using 2D convolutions. To preserve temporal information effectively in a fall video, the spatio-temporal features are extracted using a 3D CNN. In contrast to 2D CNN, a series of video frames are considered as input for 3D CNN to add a depth dimension, which can be conducted both spatially and temporally.

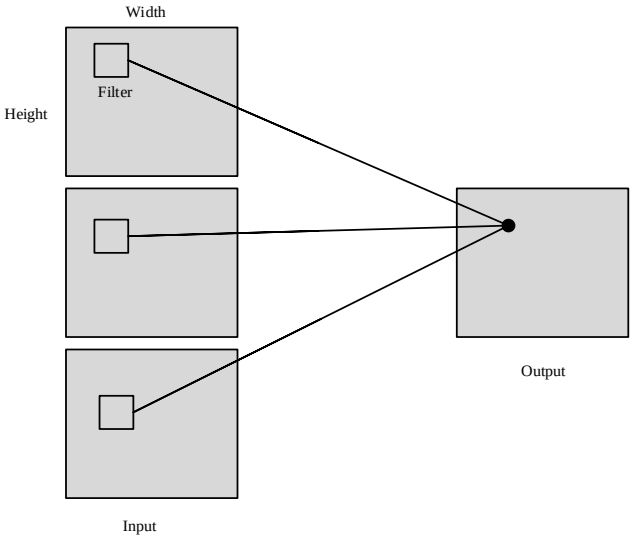


Fig. 3. 2D convolution

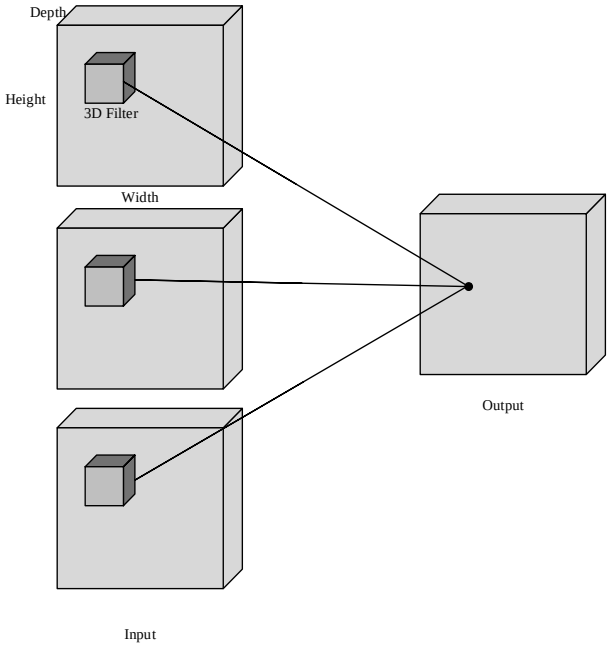


Fig. 4. 3D convolution

As shown in **Fig. 4**, feature maps of the convolution layer are related to several consecutive frames in the former layer, collecting motion information. As a consequence, three-dimensional ResNet-based method for fall detection (3D-ERes-FD) is proposed. The main component of 3D-ERes-FD is the 3D convolutional layer. Formally, the value v at position (x, y, z) is given as:

$$v_{ij}^{xyz} = f\left(\sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{s=0}^{S_i-1} w_{ijm}^{pqs} v_{(i-1)m}^{(x+p)(y+q)(z+s)}\right) + b_{ij} \quad (2)$$

in which the vertical (spatial), horizontal (spatial), and temporal extents of the filter cube w_i in i_{th} the layer can respectively be represented by P_i, Q_i, S_i. The set of feature maps from the

$(i\text{-}1)_{th}$ layer is indexed by m, and the value of the filter cube at location (p,q,s) connected to the m_{th} feature map in the previous layer can be represented. The value at location $(x+p, y+q, z+r)$ in the m_{th} feature map in the $(i\text{-}1)_{th}$ layer can be represented by $v_{(i\text{-}1)m}^{(x+p)(y+q)(z+r)}$.

3.2 The Enhanced 3D ResNet

Generally, ResNet is commonly used to deepen the network layer. The architecture of traditional residual unit is illustrated in **Fig. 5(a)**. A basic residual unit mainly includes three convolutional layers, which are respectively two $1\times1\times1$ convolutional layers and a $3\times3\times3$ convolutional layer following a batch normalization and a ReLU. And the top of the residual unit is connected to the last ReLU by a shortcut pass. In the basic residual unit, the input can be represented by x, the output can be represented by $H_t(x)$, and the residual can be represented by $F_t(x)$. Nevertheless, these $1\times1\times1$ convolutional layers that can reduce dimensionality are not useful for improving the model performance. Therefore, to improve recognition accuracy, the enhanced residual unit is designed to automatically obtain optimal features from video frames. In enhanced residual unit, a convolutional layer is added to improve the ability of feature extraction. The difference between traditional residual unit and enhanced residual unit is that the enhanced residual unit adopts two middle $3\times3\times3$ convolutional layers.

In the proposed 3D-ERes-FD, the most essential unit is enhanced residual unit, and the architecture of enhanced residual unit is illustrated in **Fig. 5(b)**. A basic enhanced residual unit mainly includes four convolutional layers, which are respectively two $1\times1\times1$ convolutional layers and two $3\times3\times3$ convolutional layers following a batch normalization and a ReLU. The capacity to extract features can be improved by increasing the $3\times3\times3$ convolutional layer. To be specific, the enhanced residual unit used in 3D-ERes-FD is computed by:

$$F_e\left(x,\{W_i\}\right)=W_4\cdot\sigma\left(W_3\cdot\sigma\left(W_2\cdot\sigma\left(W_1\cdot x\right)\right)\right) \tag{3}$$

in which the input of the residual unit can be represented by x, the weight in the i_{th} layer can be represented by W_i, the residual is represented by $F_e(x)$, and the ReLU function σ is computed by:

$$\sigma(x)=\max(0,x) \tag{4}$$

And then the input *x* is added through a shortcut, and finally output can be obtained as:

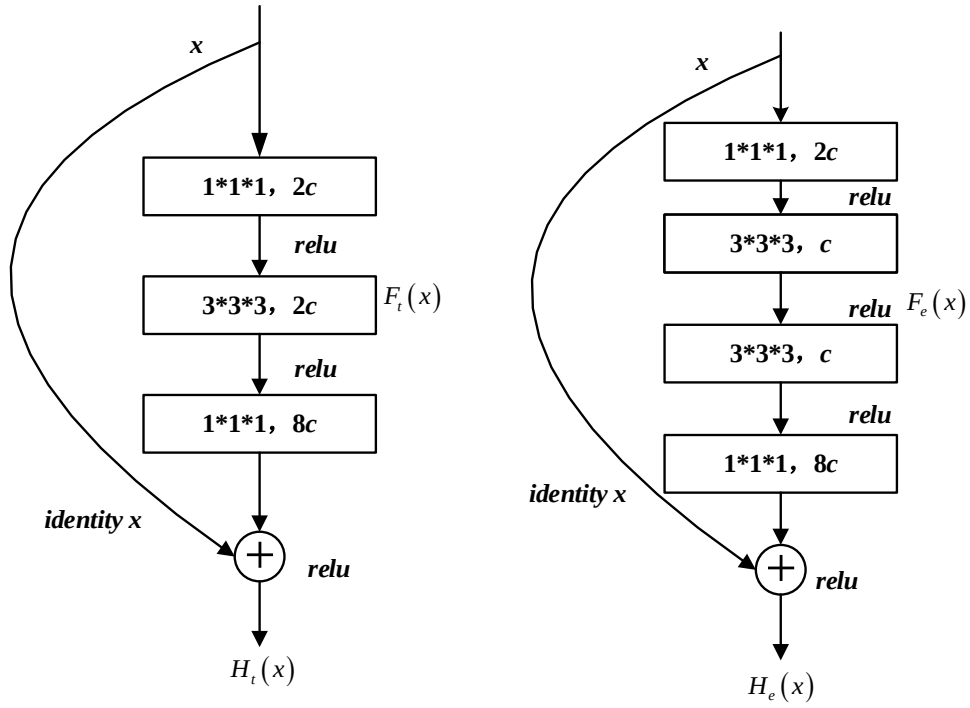
$$y=F_e\left(x,\{W_i\}\right)+W_s\cdot x \tag{5}$$

where W_s is the weight for changing the number of channels.

Additionally, after adding a $3\times3\times3$ convolutional layer, model parameters in enhanced residual unit must be larger than that in the traditional residual unit. Accordingly, to limit the number of model parameters as much as feasible, the number of channels in two $3\times3\times3$ convolutional layers are set to half of the preceding convolutional layer. As shown in **Fig. 5**, c is a constant that indicates the number of channels in residual unit. In each layer, the number of parameters can be obtained by

$$N_{output}=k\times k\times k\times p\times q \tag{6}$$

in which k is the size of convolutional kernel, p and q are respectively the number of input channels and output channels.



(a) (b)

Fig. 5. Residual unit and enhanced residual unit

To compare the number of parameters intuitively, $2c$ is assumed as the number of input channels in the residual unit. In each convolutional layer of traditional residual unit, the number of parameters can be calculated, respectively

$$N_{t1} = 1 \times 1 \times 1 \times 2c \times 2c \tag{7}$$

$$N_{t2} = 3 \times 3 \times 3 \times 2c \times 2c \tag{8}$$

$$N_{t3} = 1 \times 1 \times 1 \times 2c \times 8c \tag{9}$$

$$N_t = N_{t1} + N_{t2} + N_{t3} = 128c^2 \tag{10}$$

where $N_t = 128c^2$ represents the number of parameters in the traditional residual unit. In each convolutional layer of enhanced residual unit, the number of parameters can be calculated, respectively

$$N_{e1} = 1 \times 1 \times 1 \times 2c \times 2c \tag{11}$$

$$N_{e2} = 3 \times 3 \times 3 \times 2c \times c \tag{12}$$

$$N_{e3} = 3 \times 3 \times 3 \times c \times c \tag{13}$$

$$N_{e4} = 1 \times 1 \times 1 \times c \times 8c \tag{14}$$

$$N_e = N_{e1} + N_{e2} + N_{e3} + N_{e4} = 93c^2 \tag{15}$$

where $N_e = 93c^2$ represents the number of parameters in the enhanced residual unit. Hence, the enhanced residual unit contains fewer parameters than the traditional residual unit, which is useful to reduce the network complexity.

The proposed architecture of enhanced 3D ResNet is shown in **Table 1**, in which s in brackets indicates the stride. The input of the enhanced 3D ResNet is a six-frame RGB video clips. An image cube comprised of key frames split from a video sequence is input into the enhanced 3D-ResNet. The enhanced 3D ResNet consists of 5 residual units, including Res1, Res2, Res3, Res4 and Res5. In order to perform the downsampling, the step of the convolution kernel will be set to 2 in Res2_1, Res3_1, Res4_1 and Res5_1.

Table 1. Architecture of enhanced 3D ResNet

Layer name	Architecture
Res1	$7\times7\times7,6$
Pooling	$3\times3\times3,\ 1\times2\times2(s)$
Res2_x	$\begin{bmatrix}1\times1\times1,64\\3\times3\times3,32\\3\times3\times3,32\\1\times1\times1,256\end{bmatrix}\times3$
Res3_x	$\begin{bmatrix}1\times1\times1,128\\3\times3\times3,64\\3\times3\times3,64\\1\times1\times1,512\end{bmatrix}\times4$
Res4_x	$\begin{bmatrix}1\times1\times1,256\\3\times3\times3,128\\3\times3\times3,128\\1\times1\times1,1024\end{bmatrix}\times6$
Res5_x	$\begin{bmatrix}1\times1\times1,512\\3\times3\times3,256\\3\times3\times3,256\\1\times1\times1,2048\end{bmatrix}\times3$

For identifying a fall event, the cross-entropy loss error function J_1 is derived as:

$$J_1 = -\frac{1}{n}\sum_{i=1}^{n}\left[t_i \cdot \log p_i + (1-t_i)\log(1-p_i)\right] + \lambda\|W\|_2 \tag{16}$$

in which t_i and p_i are respectively the ground truth label and predicted classification result of the i_{th} sample, n is the total number of the samples and λ is the regularization coefficient.

4. Evaluation Experiments

All of the testing are performed on a GPU server with a 3.40 GHz Intel i7-6700 processor, 16GB of RAM, and two RTX 2080ti GPU accelerators. The proposed approach is mostly developed in Python by using the pytorch framework, and its performance is validated on several public fall datasets.

4.1 Dataset

Montreal dataset [24]: The dataset includes 24 scenarios captured with eight video cameras. Both fall and non-fall incidents are included in the first 22 scenarios. Only non-fall incidents are included in the last two scenarios. Some frames of Montreal dataset are shown in **Fig. 6**.

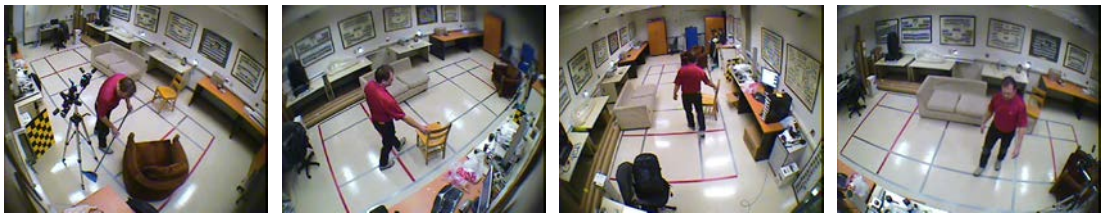


Fig. 6. Some frames of Montreal dataset

Le2i fall dataset [25]: Le2i is a RGB video set of human body actions recorded by a single camera. The video frame rate is 25 frames per second, with 320×240 pixels. Homes, coffee room, and office are among the scenarios recorded in the database, which includes 130 falls and normal activities. Some frames of Le2i fall dataset are shown in **Fig. 7**.

Fig. 7. Some frames of Le2i fall dataset

UR fall dataset [26]: We utilize the UR fall dataset from the University of Rzeszow's Computational Modelling Department in 2014. The collection sets are comprised of RGB and depth images captured by two Microsoft Kinect cameras with a resolution of 640×480 pixels, as well as accelerometer data. In this work, RGB images are utilized from camera 0 in the UR fall dataset, which contains 70 videos, respectively 30 falls and 40 normal activities. Walking, crouching, bending, and other typical daily actions are examples of non-fall frames. Fall frames mostly contain participant-performed fall actions. Some frames of UR fall dataset are shown in **Fig. 8**.

Fig. 8. Some frames of UR fall dataset

4.2 Performance Metrics

Fall detection can be usually considered as binary classification to distinguish whether or not abnormal behavior occurs. Since the probability of falling events is much lower than that of non-falling, the performance metrics of fall detection must be unaffected by imbalanced

distribution. More suitable performance to evaluate the effectiveness of such a classification method is as follow.

Precision is the percentage of precisely recognized non-fall instances among all detected non-fall instances, i.e.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

where *TP* is the number of non-fall instances accurately identified as normal, *FP* is the number of non-fall instances wrongly identified as abnormal.

Recall/sensitivity is the percentage of successfully recognized non-fall instances among all actual non-falls instances, i.e.

$$Sensitivity = \frac{TP}{TP + FN} \tag{18}$$

where *FN* is the number of fall instances wrongly classified as normal.

Specificity is the percentage of properly recognized fall instances among all actual non-fall instances:

$$Specificity = \frac{TN}{TN + FP} \tag{19}$$

where *TN* is the number of fall instances accurately identified as abnormal.

The percentage of properly recognized falls and non-fall instances can be known as accuracy:

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{20}$$

F_score is a harmonic mean of recall and precision, which is the most significant assessment metric for the overall performance of detection algorithms, i.e.

$$F\_score = 2\frac{Recall \times Precision}{Recall + Precision} \tag{21}$$

4.3 The Verification on UR Dataset

In proposed method, supervised learning is utilized, and a large number of weight parameters between the input and the output in enhanced 3D ResNet will be learnt. With minimizing the error function, optimal parameters of the enhanced 3D ResNet model can be obtained. In addition, continuous frames extracted from a fall video are taken as input to the enhanced 3D ResNet model to eliminate redundant information. Specifically, the hyper parameters are shown in **Table 2**. The dropout rate is adopted at 0.5 to avoid over-fitting, the learning rate is initially set at 0.0001, and the SGD optimizer is used to update the parameters.

Table 2. Parameter setting

Hyper parameter	Value
Learning rate	0.0001
Weight_delay	0.00003
Dropout rate	0.5
Batch_size	8
Optimizer	SGD

We further verify the advantage of the proposed 3D-ERes-FD, as compared with 3D-Res-FD which utilizes traditional residual units. The metrics of two different methods are

illustrated as **Table 3**, the model size of the proposed method is 897.5 MB, which is obviously smaller than that of the method based on 3D ResNet-50. The results can verify that the number of parameters can be reduced by adjusting the number of convolutional channels. However, in the case of deepening the network, the proposed method has a slightly longer average processing time than the method based on 3D ResNet-50, it is acceptable in a tolerable range based on the increasing depth of our network.

Table 3. Metrics of different methods on UR fall dataset

Methods	Model size	Average time/batch
3D ResNet50	897.5MB	2.73s
Proposed method	836.5MB	3.05s

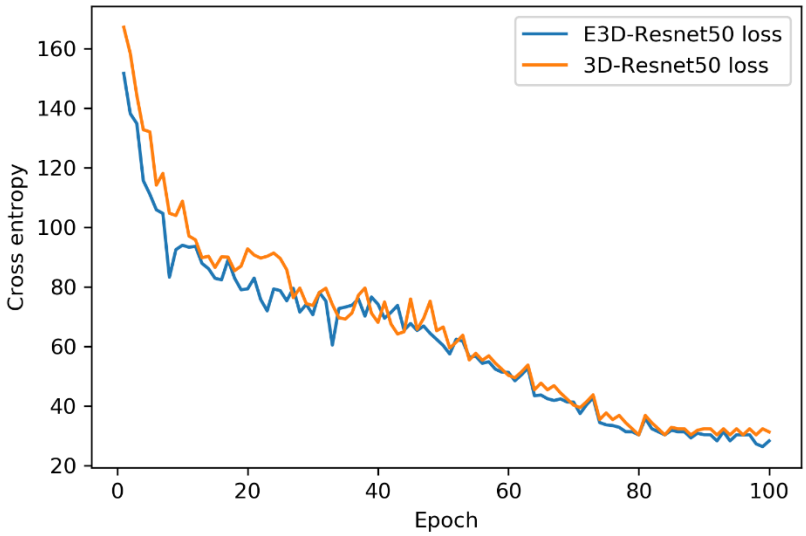


Fig. 9. The loss curve of the training process.

Additionally, **Fig. 9** displays cross entropy loss curves of two methods including 3D-ERes-FD and 3D-Res-FD in training process. As shown in **Fig. 9**, the proposed method obtains faster decay by using the enhanced residual blocks, especially in the first epoch at the beginning of training. In addition, the cross entropy loss value of 3D-ERes-FD is significantly lower than that of 3D-Res-FD, which can demonstrate that the enhanced residual block provides better learning ability. Furthermore, **Table 4** shows the model performance on the UR fall detection dataset with respect to accuracy etc. As shown in **Table 4**, performance of the proposed method is better performance than the 3D ResNet50. In particular, the accuracy is obviously increasing. These results thus confirm that, enhanced 3D ResNet can obtain spatio-temporal features with fewer parameters, and can further deepen the network to improve fall recognition accuracy.

Table 4. Comparison of different methods on UR fall dataset

Methods	Precision	Sensitivity	Specificity	Accuracy	F_score
3D ResNet50	0.95	0.904	0.926	0.914	0.926
Proposed method	1.0	0.933	1.0	**0.971**	0.965

4.4 The State-of-the-art Methods

To evaluate the effectiveness of the proposed 3D-ERes-FD, we compare the metrics, such as precision, sensitivity, specificity, accuracy and F_score, with existing methods based on the public fall detection datasets.

Montreal dataset: Kun et al. [27] create a new enhanced feature called HLC by combining HOG, LBP, and deep features, and the classification can be determined by two SVM models. Feng et al. [28] use respectively the YOLO v3 and Deep-Sort method to detect and track the pedestrians, and VGG16 is adopted to obtain the effective features, which can be classified by softmax classifier for recognizing fall event. **Table 5** illustrates the performance comparison on Montreal dataset with respect to sensitivity and specificity, it's obvious that our proposed method can achieve better performance.

Table 5. Comparison of proposed method with existing methods on Montreal dataset

Methods	Sensitivity	Specificity
Kun *et al*. [27]	0.937	0.920
Feng *et al*. [28]	0.935	0.916
Proposed method	0.947	0.990

Le2i dataset: According to features like fall angle, aspect ratio, and silhouette height, Chamle et al. [29] apply gradient boosting classifier to differentiate falls. Poonsri et al. [30] employ a mixture of Gaussian models and PCA to calculate the orientation, aspect ratio and area ratio for determining falls. Vishnu et al. [31] employ 3D-CNN to model both the appearance and motion simultaneously for obtaining effective features, and these features of conv5 layers can be categorized by polynomial support vector machine. Due to the residual unit, the network of our proposed method is deeper, and the performance will be better. As shown in **Table 6**, comparison of the proposed method with two existing computer vision-based methods [29], [30], [31] on Le2i dataset is given, it's obvious that our proposed method can achieve better performance.

Table 6. Comparison of proposed method with existing methods on Le2i dataset

Methods	Precision	Sensitivity	F_score
Chamle *et al*. [29]	0.794	0.843	0.818
Poonsri *et al*. [30]	0.891	0.931	0.911
Vishnu et al. [31]	0.815	0.930	0.868
Proposed method	0.916	0.937	0.926

UR dataset: In order to identify the human fall incidents, existing computer vision-based algorithms explore several aspects of human motions in the UR dataset. Yun et al. [32] focus on the analysis of human shapes by computing occupancy regions around the body's gravity center and extracting their angles, these features based on human shapes can be classified by SVM to distinguish fall events. Harrou et al. [33] define five lines from the silhouette's center of gravity to obtain five partial areas of human body, and combine the MEWMA charting statistic and SVM to discriminate fall events. Feng et al. [28] use respectively the YOLO v3 and Deep-Sort method to detect and track the pedestrians, and VGG16 is adopted to obtain the effective features, which can be classified by softmax classifier for recognizing fall event. Li et al. [34] adopt an unsupervised model based on auto-encoder, including three convolution layers, three deconvolution layers and LSTM, and the reconstructed error can be adopted to compute fall score to recognize fall events. In [35], the Mask-RCNN is used to extract the

human body contour from each detected binary picture, and the output of each binary image's final convolutional layer of the VGG16 is fed into the attention-guided Bi-directional LSTM model for detecting fall events. The accuracy, F score, precision, and specificity of the proposed method on the UR dataset are clearly greater than those of other methods, as shown in **Table 7**, which can fully prove the effectiveness of the proposed algorithm.

Table 7. Comparison of proposed method with existing methods on UR fall dataset

Methods	Precision	Sensitivity	Specificity	Accuracy	F_score
Yun et al. [32]	0.830	0.980	0.894	0.940	0.900
Harrou et al. [33]	0.936	1.0	0.949	0.966	0.952
Feng et al. [28]	94.8	0.914			0.931
Li et al. [34]	0.897	0.913	0.974	0.958	0.947
Chen et al. [35]	1.0	0.918	1.0	0.967	0.948
Proposed method	1.0	0.933	1.0	**0.971**	**0.965**

As validated in the experiments, our proposed method can further improve recognition accuracy on three public datasets. Compared with these methods based on traditional features, due to the ability for extracting features automatically, our proposed method is clearly more advantageous with respect to precision, specificity, accuracy and F_score. Furthermore, due to enhanced residual units with four convolutional layers, the number of parameters can be effectively reduced and the depth of the network can be deepened, thus, our proposed method can achieve better performance than those methods based on 3D-CNN. Additionally, compared with the methods based on LSTM, due to unique 3D spatial structure in proposed enhanced residual units, the proposed method is more suitable to obtain spatio-temporal features from several video frames than most methods based on LSTM, which can be further proved in the experiments.

5. Conclusion

This paper proposes a unique 50-layer 3D ResNet network and enhanced residual unit-based residual unit for fall detection. The 50-layer 3D enhanced ResNet is used as an extractor to deepen the network and improve fall identification accuracy. Additionally, it is recommended that the enhanced residual unit be used to reduce the number of convolutional channels and parameters. The experimental results show that the 3D-ERes-FD can successfully achieve accurate detection performance when compared to other advanced methods.

Acknowledgement

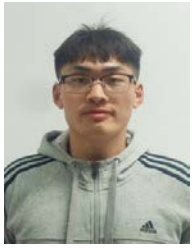
This work was supported in part by the National Nature Science Foundation of China under Grant 61473066 and Grant 61601109, and in part by the Natural Science Foundation of Hebei Province under Grant F2021501020.

References

[1] Y. Hsu, J. Perng and H. Liu, “Development of a vision based pedestrian fall detection system with back propagation neural network,” in *Proc. of 2015 IEEE/SICE International Symposium on System Integration (SII)*, Nagoya, Japan, pp. 433-437, 2015. Article (CrossRef link)

[2] L. Yang, Y. Ren, H. Hu, and B. Tian, "New fast fall detection method based on spatio-temporal context tracking of head by using depth images," *Sensors*, vol. 15, no. 9, pp. 23004-23019, Sep. 2015. Article (CrossRef link)
[3] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, no.1, pp. 144-152, Jan. 2013. Article (CrossRef link)
[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, Jul. 2006. Article (CrossRef link)
[5] A. Shahzad and K. Kim, "FallDroid: An Automated Smart-Phone-Based Fall Detection System Using Multiple Kernel Learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 35-44, Jan. 2019. Article (CrossRef link)
[6] T. Tamura, T. Yoshimura, M. Sekine, M. Uchida and O. Tanaka, "A Wearable Airbag to Prevent Fall Injuries," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 910-914, Nov. 2009. Article (CrossRef link)
[7] X. Xi, M. Tang, S. M. Miran, and Z. Luo, "Evaluation of feature extraction and recognition for activity monitoring and fall detection based on wearable sEMG sensors," *Sensors*, vol. 17, no. 6, pp. 1229-1249, May. 2017. Article (CrossRef link)
[8] D. Droghini, E. Principi, S. Squartini, P. Olivetti, and F. Piazza, "Human fall detection by using an innovative floor acoustic sensor," in *Multidisciplinary Approaches to Neural Computing, Smart Innovation, Systems and Technologies*, vol. 69, Berlin, GER, 2018, pp. 97–107. Article (CrossRef link)
[9] A. Irtaza, S. M. Adnan, S. Aziz, A. Javed, M. O. Ullah and M. T. Mahmood, "A framework for fall detection of elderly people by analyzing environmental sounds through acoustic local ternary patterns," in *Proc. of 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, AB, Canada, pp. 1558-1563, 2017. Article (CrossRef link)
[10] K. Chaccour, R. Darazi, A. Hajjam el Hassans and E. Andres, "Smart carpet using differential piezoresistive pressure sensors for elderly fall detection," in *Proc. of 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Abu Dhabi, United Arab Emirates, pp. 225-229, 2015. Article (CrossRef link)
[11] L. Ren and Y. Peng, "Research of Fall Detection and Fall Prevention Technologies: A Systematic Review," *IEEE Access*, vol. 7, no .1, pp. 77702-77722, Jun. 2019. Article (CrossRef link)
[12] Z. Bian, J. Hou, L. Chau and N. Magnenat-Thalmann, "Fall Detection Based on Body Part Tracking Using a Depth Camera," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 430-439, March 2015. Article (CrossRef link)
[13] M. Yu, S. M. Naqvi and J. Chambers, "Fall detection in the elderly by head tracking," in *Proc. of 2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, Cardiff, UK, pp. 357-360, 2009. Article (CrossRef link)
[14] D. Anderson, J. M. Keller, M. Skubic, X. Chen and Z. He, "Recognizing Falls from Silhouettes," in *Proc. of 2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, USA, pp. 6388-6391, 2006. Article (CrossRef link)
[15] R. Cucchiara, A. Prati, R. Vezzani, "A multi-camera vision system for fall detection and alarm generation," *Expert Systems*, vol. 24, no. 5, pp. 334-345, Nov. 2007. Article (CrossRef link)
[16] W. Min, L. Yao, Z. Lin, and L. Liu, ''Support vector machine approach to fall recognition based on simplified expression of human skeleton action and fast detection of start key frame using torso angle," *IET Computer Vision*, vol. 12, no. 8, pp. 1133-1140, Dec. 2018. Article (CrossRef link)
[17] W. Lie, A. T. Le and G. Lin, "Human fall-down event detection based on 2D skeletons and deep learning approach," in *Proc. of 2018 International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Thailand, pp. 1-4, 2018. Article (CrossRef link)
[18] X. Cai, S. Li, X. Liu and G. Han, "A Novel Method Based on Optical Flow Combining with Wide Residual Network for Fall Detection," in *Proc. of 2019 IEEE 19th International Conference on Communication Technology (ICCT)*, Xi'an, China, pp. 715-718, 2019. Article (CrossRef link)
[19] A. Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless Communication and Mobile Computing*, vol. 2017, no. 1, pp. 1-16, Dec. 2017. Article (CrossRef link)

[20] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, no. 1, pp. 32-43, Jul. 2018. Article (CrossRef link)
[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013. Article (CrossRef link)
[22] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 314-323, Jan. 2019. Article (CrossRef link)
[23] Y. Wang, W. Zhou, Q. Zhang, and H. Li, "Enhanced action recognition with visual attribute-augmented 3D convolutional neural network," in *Proc. of 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, San Diego, CA, USA, pp. 1-4, 2018. Article (CrossRef link)
[24] M. Li, G. Xu, B. He, X. Ma and J. Xie, "Pre-Impact Fall Detection Based on a Modified Zero Moment Point Criterion Using Data from Kinect Sensors," *IEEE Sensors Journal*, vol. 18, no. 13, pp. 5522-5531, July. 2018. Article (CrossRef link)
[25] I. Charfi, J. Mitéran, J. Dubois, M. Atri and R. Tourki, "Optimised spatio-temporal descriptors for real-time fall detection: Comparison of SVM and Adaboost based classification," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 041106-041106, Oct. 2013. Article (CrossRef link)
[26] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Comput. Methods Programs Biomed*, vol. 117, no. 3, pp. 489–501, Dec. 2014. Article (CrossRef link)
[27] K. Wang, G. Cao, D. Meng, W. Chen, and W. Cao, "Automatic fall detection of human in video using combination of features," in *Proc. of 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, China, pp. 1228–1233, 2016. Article (CrossRef link)
[28] Q. Feng, C. Gao, L. Wang, Y. Zhao, T. Song, and Q. Li, "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," *Pattern Recognition Letter*, vol. 130, no. 1, pp. 242–249, Feb. 2020. Article (CrossRef link)
[29] M. Chamle, K. G. Gunale and K. K. Warhade, "Automated unusual event detection in video surveillance," in *Proc. of 2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp. 1-4, 2016. Article (CrossRef link)
[30] A. Poonsri and W. Chiracharit, "Improvement of fall detection using consecutive-frame voting," in *Proc. of 2018 International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Thailand, pp. 1–4, 2018. Article (CrossRef link)
[31] C. Vishnu, R. Datla, D. Roy, S. Babu and C. K. Mohan, "Human Fall Detection in Surveillance Videos Using Fall Motion Vector Modeling," *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17162-17170, Aug. 2021. Article (CrossRef link)
[32] Y. Yun and I. Y. Gu, "Human fall detection via shape analysis on Riemannian manifolds with applications to elderly care," in *Proc. of 2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, Canada, 2015, pp. 3280-3284, 2015. Article (CrossRef link)
[33] F. Harrou, N. Zerrouki, Y. Sun and A. Houacine, "Vision-based fall detection system for improving safety of elderly people," *IEEE Instrumentation & Measurement Magazine*, vol. 20, no. 6, pp. 49-55, Dec. 2017. Article (CrossRef link)
[34] S. Li, X. Song, S. Xu, H. Qi and Y. Xue, "Dilated spatial–temporal convolutional auto-encoders for human fall detection in surveillance videos," *ICT Express*, July. 2022. Article (CrossRef link).
[35] Y. Chen, W. Li, L. Wang, J. Hu, and M. Ye, "Vision-based fall event detection in complex background using attention guided bi-directional LSTM," *IEEE Access*, vol. 8, no.1, pp. 161337-161348, Sep. 2020. Article (CrossRef link)



Suyuan Li received the B.S. degree from Liaoning Normal University, Dalian, in 2017 and the M.E. degree in electronics and communication engineering from Northeastern University, Shenyang, China in 2020. He is currently pursuing the Ph.D. degree information and communication engineering from Northeastern University, China. His research interests include fall detection based on deep learning and image processing.

Xin Song was born in Jilin, China, in 1978. She received her Ph.D. degree in Communication and Information System in Northeastern University in China in 2008. She is now a teacher working in Northeastern University at Qinhuangdao, China. Her research interests are in the area of robust adaptive beam-forming and wireless communication.

Jing Cao received the B.S. degree in Hebei University of Economics and business, Hebei, China and the M.S. degree in Yanshan University, Hebei, China. She is an Associate Professor with Hebei Normal University of Science & Technology at Qinhuangdao, China. She is currently pursuing the Ph.D. degree in Communication and Information System at Northeastern University, Shenyang, China. Her research interests include D2D communications, green communications and wireless resource allocation.

Siyang Xu received the B.E. degree in Electronics and Information Engineering from the University of Science and Technology Liaoning, China, 2016 and the M.S. degree in Computer Science and Engineering department from Northeastern University (NEU), Shenyang, China in 2018. He is currently pursuing the Ph.D. degree in Communication and Information System with Northeastern University (NEU), Shenyang, China. His current research interests include energy harvesting, physical layer security and relay cooperative communication.