Convolutional GRU and Attention based Fall Detection Integrating with Human Body Keypoints and DensePose

Yi Zheng[1], Cunyi Liao[2], Ruifeng Xiao[1], and Qiang He[1*]
[1] Intelligent Medical Engineering Research Center, School of Artificial Intelligence, Jianghan University
Wuhan, 430056 China
[2] College of Physical Science and Technology, Central China Normal University
Wuhan, 430079 China
[e-mail: qh2020@jhun.edu.cn*]
[*]Corresponding author: Qiang He

Received April 11, 2024; revised July 29, 2024; accepted August 30, 2024; published September 30, 2024

Abstract

The integration of artificial intelligence technology with medicine has rapidly evolved, with increasing demands for quality of life. However, falls remain a significant risk leading to severe injuries and fatalities, especially among the elderly. Therefore, the development and application of computer vision-based fall detection technologies have become increasingly important. In this paper, firstly, the keypoint detection algorithm ViTPose++ is used to obtain the coordinates of human body keypoints from the camera images. Human skeletal feature maps are generated from this keypoint coordinate information. Meanwhile, human dense feature maps are produced based on the DensePose algorithm. Then, these two types of feature maps are confused as dual-channel inputs for the model. The convolutional gated recurrent unit is introduced to extract the frame-to-frame relevance in the process of falling. To further integrate features across three dimensions (spatio-temporal-channel), a dual-channel fall detection algorithm based on video streams is proposed by combining the Convolutional Block Attention Module (CBAM) with the ConvGRU. Finally, experiments on the public UR Fall Detection Dataset demonstrate that the improved ConvGRU-CBAM achieves an F1 score of 92.86% and an AUC of 95.34%.

Keywords: Fall detection, ConvGRU, CBAM, Skeletal keypoint detection, DensePose

This work is supported in part by The Research Fund of Jianghan University (Grant No. 2022SXZX29), and in part by Hubei Province health and family planning scientific research project (Grant No. WJ2023F039).

http://doi.org/10.3837/tiis.2024.09.016 **ISSN : 1976-7277**

1. Introduction

With the intensification of aging, falls have become the foremost factor affecting the health and later life of the elderly, being one of the primary causes of fractures, brain injuries, and other bodily harm. Furthermore, falls constitute a severe public health issue, imposing substantial burdens on the healthcare and social welfare systems, including medical expenses, rehabilitation costs, and potentially long-term care costs. At the same time, fall detection technology also offers crucial support for the independent living of individuals with disabilities, making the work of fall detection incredibly vital [1].

Currently, there has been a significant body of representative research in fall detection, primarily divided into three types: systems based on wearable sensor devices [2-5], environment-installed sensors [6,7], and vision technology [8-10]. Wearable systems typically measure the subject's movement through accelerometers worn on the body [2,3] and identify falls using algorithms [4,5], but they can interfere with normal activities. Meanwhile, sensor devices can be installed within the environment [6], such as on walls [7], but this approach is subject to environmental constraints and high costs. At present, the most widely applied method involves using cameras [8,10] combined with deep learning algorithms for fall detection, although this can lead to significant prediction time costs. These camera-based fall detection systems rely heavily on real-time data transmission and processing, and robust and efficient 5G wireless communication technologies now provide reliable and efficient data transmission guarantees. Secondly, the multi-area camera, by means of edge computing, allows the deployment of multi-camera systems to perform initial data processing and analysis on edge devices, and then transmit the results of the processing to a centralized server or cloud service via a wireless network. Finally through wireless communication technology, the camera-based fall detection system can be remotely monitored and supported. Healthcare professionals can access the patient's activity status in real time over the wireless network and quickly intervene before a fall event occurs.

To address the challenge of predicting falls before they occur, we consider the use of dual-channel feature maps. Skeletal keypoint maps and dense feature maps enable an accurate understanding and analysis of human posture and movement. Therefore, this paper utilizes dual-channel information as the foundation for our research. Initially, the ViTPose++ keypoint detection algorithm [11] is applied to the public UR Fall Detection Dataset [12] for keypoint detection, thereby acquiring skeletal keypoint data to generate skeletal feature maps. Subsequently, the DensePose algorithm [13] is used on the same dataset to produce dense feature maps of the human body, thereby enriching the data features through a dual-channel approach.

In this study, we propose a fall detection algorithm that leverages the strengths of Convolutional Gated Recurrent Unit (ConvGRU) and Convolutional Block Attention Module (CBAM). The choice of these algorithms is driven by their respective advantages in handling spatio-temporal data and enhancing feature representation.

ConvGRU [66] combines the capabilities of Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU), allowing for effective extraction of spatial features from image sequences and capturing temporal dependencies crucial for fall detection. Compared to traditional GRU, ConvGRU employs convolution operations that better capture spatial context within video frames, essential for recognizing fall events. Meanwhile, CBAM [14] integrates spatial and channel attention mechanisms, significantly improving the model's ability to focus

on important features while suppressing irrelevant background noise. This enhancement is critical in fall detection tasks, where distinguishing between relevant human movements and background activities can greatly affect accuracy.

In this paper, we designed a dual-channel fall detection model. Initially, we use the ViTPose++ algorithm to extract human keypoint coordinates and generate skeletal feature maps. Simultaneously, DensePose algorithm is used to create dense feature maps of the human body. These feature maps serve as dual-channel inputs to the ConvGRU-CBAM model. The ConvGRU processes temporal correlations, while CBAM enhances feature representation, leading to efficient and accurate fall detection. The introduction of the attention mechanism enables the model to suppress some environmental noise and irrelevant actions, ultimately achieving an F1 score of 0.9286 on a public UR Fall Detection Dataset.

2. Related work

The current fall detection systems [15-20] have seen significant achievements by scholars in various directions, mainly categorized into three types. Systems based on wearable sensors [21,22] generally perform fall detection by measuring certain kinematic parameters. For instance, Barshan et al. [23] introduced a heuristic fall detection algorithm that combines the double threshold of two simple features and fuzzy logic technology to extract features from the accelerometer and gyroscope data recorded by a motion sensor unit worn on the waist, offering good real-time performance but at the inconvenience of wearing, which affects daily activities. On the other hand, the lack of real data remains a significant challenge for fall detection algorithms. Mosquera-Lopez et al. [24] addressed this issue by proposing a context-aware fall detection system algorithm based on inertial sensors and flight time sensors, using fall data provided by patients with Multiple Sclerosis (MS). This algorithm uses an autoencoder combined with the reconstruction error of the worn accelerometer signal to detect fall candidates, followed by fall detection using a balanced random forest trained on acceleration and motion features. However, this approach has limitations, such as affecting real-time performance due to two rounds of candidate detection and increasing the inconvenience of walking for patients due to wearing sensors, which does not guarantee the accuracy of real data. Nonetheless, with the continuous development of smart healthcare, fall detection algorithms can also be applied to public healthcare systems, but there are limitations in operability and technical issues such as high power consumption. Therefore, Qian et al. [25] proposed a wearable fall detection system based on a multi-level threshold algorithm that combines Micro-Electro-Mechanical Systems (MEMS) with Narrowband Internet of Things (NB-IoT) for fall detection, providing a user interface tailored for healthcare professionals. Among these methods, each has its advantages and limitations, thus making the balance between real-time performance, data accuracy, and user convenience a focal point of research in fall detection systems.

To address the inconvenience of wearing sensors, devices can be installed within the environment, with numerous studies utilizing radar and Wi-Fi for detection. Radar sensor systems [26-28], with their non-contact nature and adaptability to environmental conditions, offer a potential solution for fall detection. For instance, He et al. [29] developed a novel non-contact fall detector based on MEMS low-resolution infrared sensors and radar sensors for detecting falls beside the bed. Additionally, a millimeter-wave Frequency-Modulated Continuous-Wave (FMCW) radar based on Pattern Contour Vector (PCV) [30] can also be used for fall detection, featuring advantages such as not requiring to be worn, being inconspicuous, non-invasive, and privacy-preserving. With the continuous advancement of

deep learning, increasing research integrates neural networks with traditional methods [32,33] to enhance the accuracy of fall detection. For example, Sadreazami et al. [31] applied short-time Fourier transform to each radar echo signal for time-frequency analysis, converting the resulting spectrograms into binary images as inputs for convolutional neural networks. Moreover, Wi-Fi technology [34-38] has also shown potential in the field of fall detection. These studies demonstrate various methods of utilizing Wi-Fi signals for fall detection but also face challenges regarding signal accuracy, environmental interference, and reliability.

With the evolution of computer vision technology, image-based fall detection systems [39,40,42] are increasingly emerging. For instance, an unsupervised method for fall detection [41] initially converts RGB video frames into human posture images to eliminate background interference, focus on human motion, and protect privacy. Then, based on conditional Generative Adversarial Networks, it uses a sequence of historical human posture images to predict future posture images. Finally, fall detection is achieved by using the prediction error of human posture images and the anomaly scores calculated from traditional handcrafted features of the actual posture. Currently, deep learning employs multimodal methods [43-45] for fall detection, but multimodal approaches face challenges in maximizing the value of information collected from different modalities during clinical assessments and in enhancing the performance of fall detection.

Image-based fall detection systems have become the most popular method due to their low cost and lack of environmental constraints. Bedside falls are a critical issue in elderly care, leading to the proposal of various bedside monitoring systems [15] for detecting such falls. Fall detection based on human keypoints is another direction in image technology, with numerous algorithms for detecting human keypoints emerging. Among these, the ViTPose++ algorithm, based on the Vision Transformer, is currently the most favored for keypoint detection. This algorithm has been widely applied in areas such as human posture estimation [47,48,67], action recognition [49], and visual object detection [50]. For instance, [51] and others have utilized keypoints obtained by the ViTPose algorithm to propose a multimodal gait recognition method, which integrates multiple posture representation models to comprehensively describe the way people walk.

Additionally, the use of dense maps from DensePose for posture analysis is another image technology application, also extensively used in human-related virtual reality [52-55], object detection segmentation [55-58], posture estimation [59-61], and action recognition tracking [62-64]. For example, [65] and others employed the DensePose human body model and posture extraction strategy to construct an industrial robot posture recognition model, then accurately and efficiently recognized the posture of industrial robots by inputting individual robot images into a high-precision posture estimation network. Therefore, this paper first generates human skeletal graphs using the ViTPose++ [11] keypoint detection algorithm, then produces dense feature maps of the human body using the DensePose [13] algorithm. Based on image vision technology, combined with deep learning's recurrent neural networks (RNN) and attention mechanisms, a dual-channel fall prediction system based on video streams is proposed, utilizing both human skeletal graphs and dense feature maps.

3. Data preparation

3.1 Human keypoint detection

The Vision Transformer (ViT) [46] is a deep learning model architecture for the field of computer vision, surpassing traditional CNN models in tasks such as image classification, semantic segmentation, and keypoint detection. ViTPose++ represents the first integration of the ViT keypoint detection algorithm in the task of posture estimation, using the ViT structure as the backbone along with a lightweight decoder, making it the most effective model for detecting keypoints on the MS COCO validation set currently.

This paper utilizes ViTPose++ to acquire a human skeletal keypoint dataset, employing the same data format as the MS COCO keypoint dataset, which includes the two-dimensional coordinates and confidence levels of 17 human body keypoints. By detecting every frame, a one-dimensional keypoint vector H' is obtained by

$$H' = [(x'_0, y'_0, s'_0), \dots, (x'_i, y'_i, s'_i)], (0 \le i \le 16) \quad (1)$$

where, x' represents the horizontal coordinate of the keypoint, y' denotes the vertical coordinate of the keypoint, and s' signifies the confidence level of the keypoint's coordinates.

The extracted human skeletal keypoints are as shown in **Fig. 1(a)**, where the 17 human body keypoints correspond to the keypoints numbered 0 to 16 in **Table 1**.

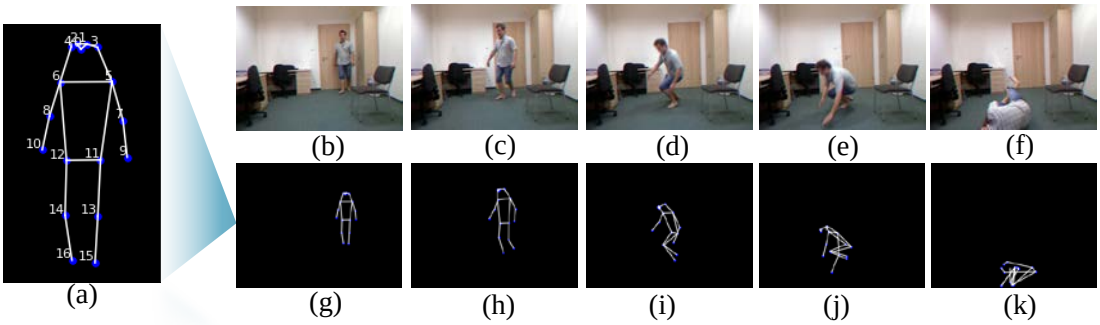


Fig. 1. The diagram of the fall process and its human skeleton, (a) displays the extracted human skeletal keypoints, (b)-(f) are the original images, (g)-(k) are the corresponding human skeleton diagrams

Using the detected keypoint coordinates, human skeletal feature maps are generated through the keypoint connection and image drawing. First, based on the prior knowledge of the human skeletal structure, the detected keypoints are connected. For example, the left shoulder is connected to the left elbow, and the left elbow is connected to the left wrist, forming the skeleton structure. Second, the image drawing library is employed to draw these connections on a blank image, creating the corresponding skeletal feature map. Each skeletal connection is represented by a line, with different colors used to distinguish between the left and right limbs and various body parts.

Fig. 1(b) to 1-k display five original images and their corresponding human skeletal graphs extracted from a fall video. After keypoints are detected in each frame of the preprocessed data sequence $Frame(t_0) \sim Frame(t_M)$, a keypoint matrix K'_M composed of M one-dimensional keypoint vectors H' is obtained as

$$K'_M = [H'_0, \dots, H'_t, \dots, H'_{M-1}]^T \quad (2)$$

where, $0 < t \leq M - 1$, K'_M represents the keypoint matrix in a video stream.

Table 1. Human body keypoint name

Number	Keypoint	Number	Keypoint
0	Nose	9	Left wrist
1	Left eye	10	Right wrist
2	Right eye	11	Left hip
3	Left ear	12	Right hip
4	Right ear	13	Left knee
5	Left shoulder	14	Right knee
6	Right shoulder	15	Left ankle
7	Left elbow	16	Right ankle
8	Right elbow	17	Head

3.2 Importance analysis

The importance analysis of Random Forest (RF) is a method used to determine the impact of features on model performance. In this paper, RF is utilized to ascertain the importance of 9 human joint features. RF evaluates feature importance by the average increase in node of Gini impurity brought by each feature across all trees in the forest. The more a feature decreases impurity, the higher its importance score. As shown in the blue bar graph in **Fig. 2**, with a threshold of I=0.02, features below this threshold, such as Ear, Eye, and Nose, are considered less important. Therefore, the central points of these three feature joints are combined to form a single feature joint named Head, ultimately consolidating into 7 human joint features. Subsequently, Random Forest is used again to obtain the feature importance, as illustrated in the orange bar graph in **Fig. 2**, with the corresponding 13 human keypoints being those numbered 5 to 17 in **Table 1**.

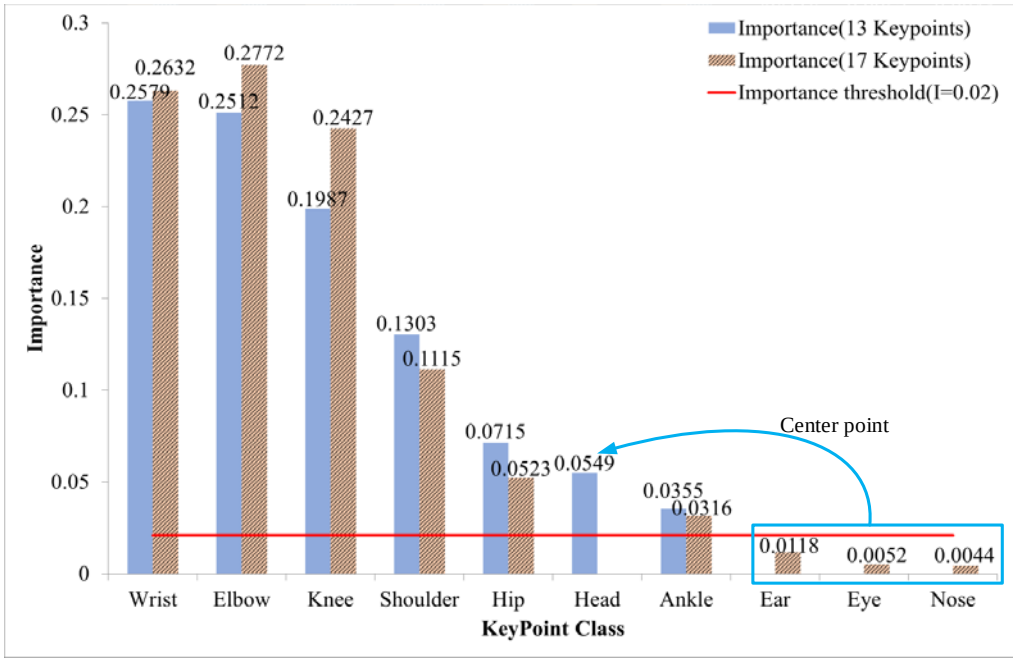


Fig. 2. The importance of each body keypoints

Following the analysis of feature importance, a new keypoint matrix is generated based on the updated information.

$$K_M = [H_0, \dots, H_t, \dots, H_{M-1}]^T \tag{3}$$

where, H represents the updated one-dimensional keypoint vector, which is defined as:

$$H = [(x_0, y_0, s_0), \dots, (x_i, y_i, s_i)], (0 \le i \le 12) \tag{4}$$

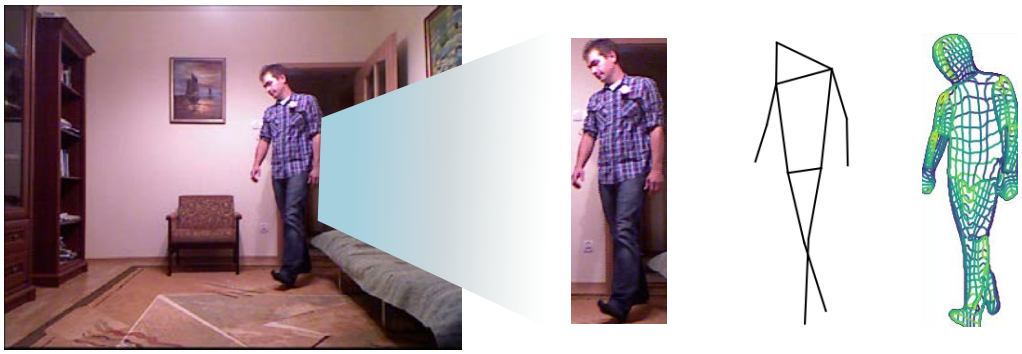
where, x_i represents the horizontal coordinate of the keypoint, y_i denotes the vertical coordinate of the keypoint, and s_i signifies the confidence level of the keypoint's coordinates.

3.3 Dual-channel feature maps

The necessity for a dual-channel design primarily stems from the absence of skeletal and human keypoint information. In scenarios where skeletal data is missing, traditional single-channel models may fail to capture comprehensive posture information. Similarly, when human keypoints are missing, such as in cases of partial occlusion, single-channel models can lose accurate comprehension of the body structure. To compensate for these information gaps, dense feature maps from DensePose are introduced. By employing a dual-channel structure, both posture and keypoint information are captured more comprehensively, thereby enhancing the model's ability to interpret complex scenes and ensuring more accurate and detailed human posture analysis. This design emphasizes not only the accuracy of keypoints but also the overall coherence of posture, providing richer and more precise inputs for the task of fall detection.

3.3.1 Human skeletal feature maps

Human skeletal feature maps are generated using one-dimensional keypoint vectors , and by cropping according to the size of the bounding box, the skeletal part of the body is extracted to obtain the final human skeletal feature map, as shown in **Fig. 3**. In this context, **Fig. 3(a)** represents the original image, and **Fig. 3(b)** depicts the human target detection box image.



(a) (b) (c) (d)

Fig. 3. Dual-channel feature maps: (a) Original figure. (b) Human bounding box. (c) Human Skeletal feature maps. (d) Human compact feature maps.

3.3.2 Human compact feature maps

Dense feature maps are acquired from the DensePose algorithm [13] and crops them according to the size of the bounding box to extract the human body part, resulting in the final dense feature map of the human body, as shown in **Fig. 3(d)**.

4. Algorithm

Since falling is a continuous process, the human posture at the current moment is significantly related to the posture at the previous moment. Therefore, this paper utilizes ConvGRU for fall prediction and employs CBAM for feature fusion on this basis, with the overall algorithm process depicted in **Fig. 4**.

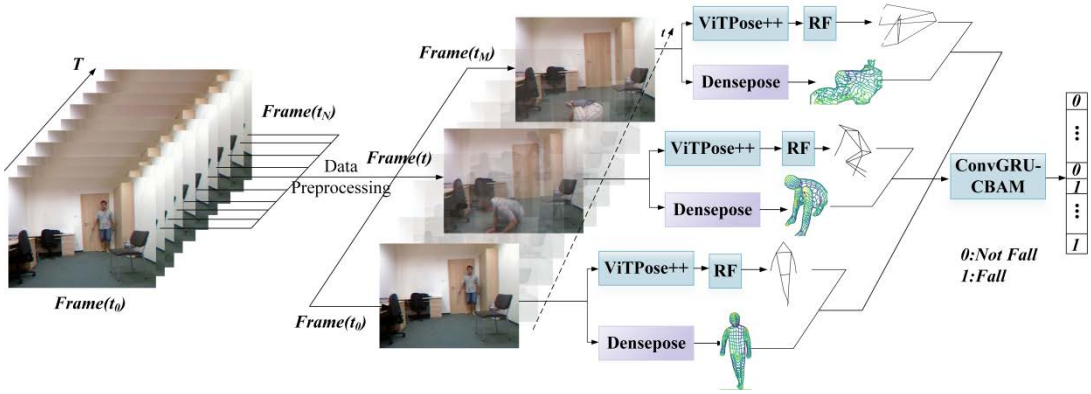


Fig. 4. The overall block diagram of human keypoint-guided fall detection algorithm

(1) After preprocessing the video sequence, the ViTPose++ keypoint detection algorithm is used to detect the coordinates and confidence levels of 17 human keypoints for each frame, resulting in a corresponding one-dimensional keypoint vector *H*. After processing the entire video sequence, a keypoint matrix K_M composed of M one-dimensional vectors H is obtained.
(2) Feature importance analysis is conducted using Random Forest (RF) to filter out features with lower contributions to the fall prediction task and update the keypoint matrix K'_M. Subsequently, human skeletal feature maps are generated based on the keypoint matrix, and the human body part is cropped according to its target detection box.
(3) The DensePose algorithm is utilized to generate dense feature maps of the human body, and the human body part is cropped based on its target detection box.
(4) The obtained human skeletal feature maps and corresponding dense feature maps are used as dual-channel inputs for the ConvGRU-CBAM model, and the model is trained for the fall detection task.

4.1 ConvGRU-CBAM

GRU, a variant within the Recurrent Neural Network (RNN) family, addresses the issues of vanishing and exploding gradients common in traditional RNNs through its gating mechanism. It offers a simpler structure and fewer parameters than the Long Short-Term Memory (LSTM) network, yet maintains comparable performance. Compared to GRU, the Convolutional Gated Recurrent Unit (ConvGRU) neural network exhibits stronger learning capabilities, making it the primary model for fall prediction tasks in this paper. Additionally, the Attention Mechanism, widely used in machine learning and deep learning, simulates

human attention and memory capabilities in processing information, suitable for handling sequential data and achieving sequence-to-sequence mapping. To enhance feature integration, we employ the Convolutional Block Attention Module (CBAM), combining spatial and channel attention mechanisms within the ConvGRU model to boost performance. Against this backdrop, the model addresses the extraction of fall-related features across three dimensions: temporal (the relative displacement tracking in time sequences by ConvGRU), spatial (the relative positional relationships of human keypoints), and channel (the relationships between the dual input feature maps), integrating these dimensions through CBAM. Experimental results also demonstrate an improvement in model performance by extracting fall features across these three dimensions (temporal, spatial, and channel) and incorporating the CBAM attention mechanism. The ConvGRU-CBAM network structure is illustrated in **Fig. 5**.

After obtaining the human key feature maps and dense feature maps from a video sequence through the ViTPose++ and DensePose algorithms, respectively, these two types of feature maps are used as dual-channel inputs for the ConvGRU-CBAM model. Similar to GRU, the ConvGRU within the model also contains a reset gate and an update gate, with the gating signals represented in **Fig. 5** as r_t and z_t, respectively.

$$r_t = \sigma_s(W_r(h_{t-1} \oplus x_t)) \tag{5}$$

$$z_t = \sigma_s(W_z(h_{t-1} \oplus x_t)) \tag{6}$$

where W_r and W_z represent corresponding tensors, h_{t-1} denotes the state from the previous moment, and x_t indicates the input at the current moment, which refers to the dual-channel feature map at the current time. The symbol * represents the convolution operation, and σ_s signifies the sigmoid function, capable of transforming data into a range between 0 and 1, thereby acting as a gating signal. The reset gate r_t is used to select and forget a portion of the fall information extracted from the past, while the update gate z_t decides how much of the previously extracted fall-related information should be copied and passed on to the next moment, in order to focus on the transition of falling postures between sequences.

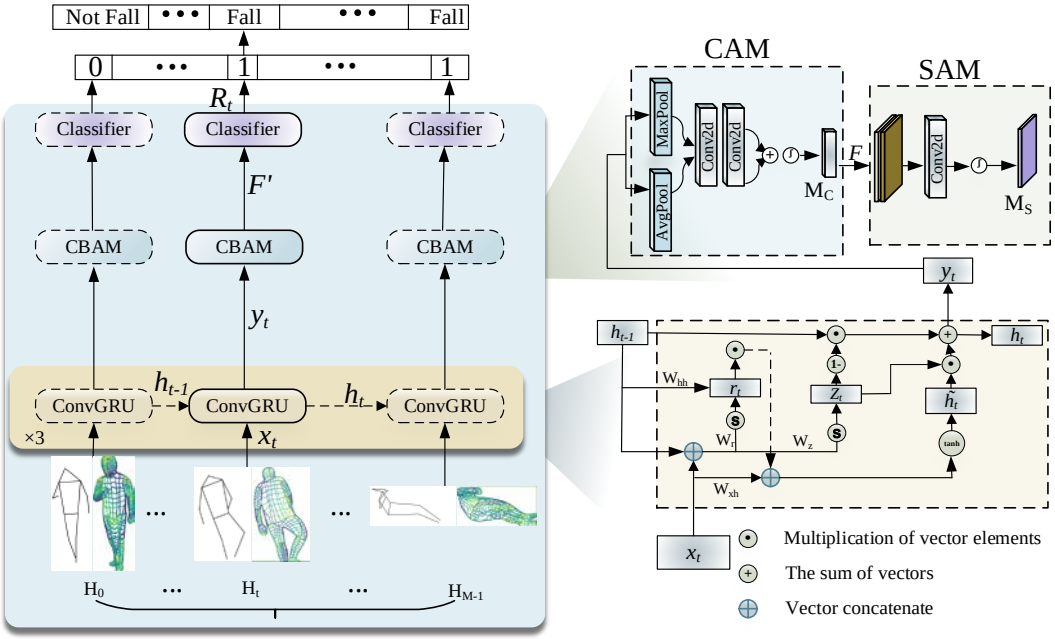


Fig. 5. ConvGRU-CBAM model diagram

The reset gate and the activation function Relu are used to process the state from the previous moment and the dual-channel feature map input x_t at the current moment, as follows,

$$h'_{t-1} = h_{t-1} \odot r_t \tag{7}$$

$$\tilde{h}_t = \sigma_t\left(W(h'_{t-1} \oplus x_t)\right) \tag{8}$$

where, $\odot$ denotes the Hadamard product operation. The updated value $\tilde{h}_t$ is obtained through the activation function ReLu σ_t .

After passing the updated value through the update gate for selection, it is added to the selected output information of the hidden state from the previous moment, resulting in the output h_t and hidden layer output state y_t for the current moment, which is expressed as:

$$y_t = h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{9}$$

After the gating mechanism of the first ConvGRU layer makes its selection, the output h_t at the current moment is used as the input for the next moment. Meanwhile, the output state y_t of the hidden layer serves as the input for the second and third ConvGRU layers. Following the same process of selection through the reset gate and update gate, the output y'_t is then used as the input for the CBAM module.

CBAM consists of the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), as shown in the CBAM block in **Fig. 5**. The CAM is designed to capture the relationships between different channels of the feature map, adaptively weighting the feature matrix by calculating the importance of each channel to enhance the interaction and combination capabilities among different channels. In the CAM section, spatial dimension reduction of y'_t is first achieved through average pooling, while max pooling is used to infer attention on finer channels. Then, channel attention feature weights M_c are generated through two consecutive convolution operations,

$$M_c = \sigma_s(\mathcal{F}_{conv}(\mathcal{F}_{avgpool}(y'_t)) + \mathcal{F}_{conv}(\mathcal{F}_{maxpool}(y'_t))) \tag{10}$$

where, $\mathcal{F}_{conv}$ denotes the two consecutive convolution operations. $\mathcal{F}_{avgpool}$ represents average pooling, while $\mathcal{F}_{maxpool}$ stands for max pooling. The channel attention is then obtained by multiplying the result with M_c and the input y'_t of the CAM section. represents element-wise multiplication.

$$F = M_c \otimes y'_t \tag{11}$$

The Spatial Attention Module (SAM) aims to capture the relationships between different spatial positions within the feature map by analyzing the spatial distribution information of the feature map to determine the importance of different positions. This enhances the network's perception of spatial structures. In the SAM section, channel attention F is processed in the same manner through average pooling and max pooling operations along the channel dimension. The outputs of these operations are then concatenated into a single feature descriptor, which is further processed by a convolutional layer to generate spatial attention feature weights M_s.

$$M_s = \sigma_s(\mathcal{F}_{conv}(\mathcal{F}_{avgpool}(F) + \mathcal{F}_{conv}(\mathcal{F}_{maxpool}(F)))) \tag{12}$$

where, the spatial attention feature weights M_s are then multiplied by the input F of the SAM section to obtain the final output F of the CBAM module, which is,

$$F' = M_s \otimes F \tag{13}$$

where, F'represents the weights or attention levels of each element in the input sequence $y_t^{'}$. After processing this weight file F' through a classifier, the output R_t of the model at the current moment is obtained, which is the status judgment of the current frame. A value of 0 indicates that the person is not in a lying down state, and 1 indicates that the person is in a lying down state. In case that after analyzing the entire video sequence K_M, a rising edge result is obtained in $[0,0,\ldots,\ldots,1,1]$, it is determined to be a fall.

4.2 Training strategies

In the training task of this paper, the MSE loss is used to guide the adjustment of model parameters during the training process, and the Adam optimizer is employed to perform parameter updates. The performance of the model is enhanced by minimizing the loss function. Through the adjustment of hyperparameters, the model is optimized, and the hyperparameters are shown in **Table 2**.

The experiments demonstrate that setting the initial learning rate to 0.0001 enables the model to achieve optimal performance. Additionally, the model is set to train for 250 epochs, but to prevent overfitting, an early stopping strategy is employed to enhance the model's efficiency and generalization performance. Specifically, training is halted if there is no significant improvement on the validation set after 20 consecutive epochs. When validated on the test set, since the model outputs the status of each frame in the video sequence and determines whether a fall has occurred based on the entire sequence's status.

Table 2. Hyperparameter during the training step

Parameter	Data
learning rate	0.0001
Patience	20
Batch size	8
Epochs	250

4.3 Model deployment

The trained model is deployed on the Azure Kinect depth camera, where human keypoints are extracted as inputs to the deployed model for fall detection. The specific implementation flow is illustrated in **Fig. 6**.

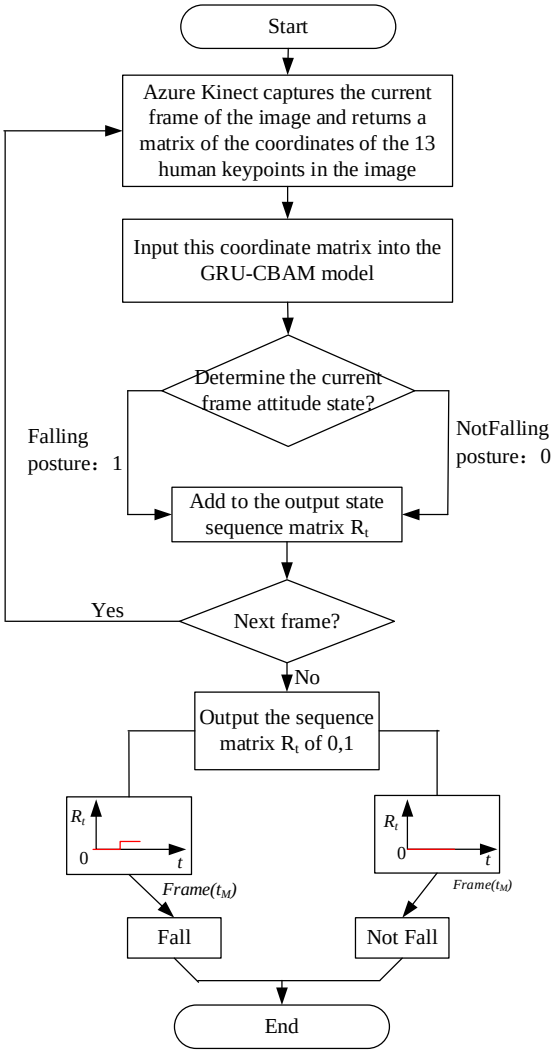


Fig. 6. Deployment flowchart

(1) The Azure Kinect depth camera captures the current frame image, identifies the 13 human body keypoints corresponding to numbers 5 to 17 in **Table 1**, and returns their coordinates to form a keypoint coordinate matrix, while also generating a human skeletal feature map.

(2) The current frame image is captured from the camera, and a dense feature map of the human body is generated based on the DensePose algorithm.

(3) The two types of feature maps obtained above are input into the deployed ConvGRU-CBAM model in a dual-channel manner to determine the posture of the current frame.

(4) The returned result is added to the output status sequence matrix, the process continues with capturing the next frame, repeating steps (1) and (2) until the end of the video.

(5) After outputting a complete sequence matrix, the presence of a fall is determined based on whether a rising edge appears in the sequence. If a rising edge exists, it is determined to be a fall. If it is a straight line of zeros, there is no fall.

5. Experiment results

5.1 Dataset

Due to privacy concerns making it difficult to collect data, there are not many available datasets for fall detection, and they are mostly in the form of RGB videos. This paper extracts human skeletal keypoint datasets from these datasets using keypoint detection algorithms for fall detection.

The UR Fall Detection Dataset [12] contains 30 fall sequences (Fall) and 40 daily life activity sequences (Adl), with each video sequence having a frame rate of 30 FPS and a resolution of 640×480.

After obtaining human skeletal feature maps and dense feature maps using the UR Fall Detection Dataset, each original sequence in the dataset is split into several shorter sequences by skipping frames, with each short sequence having around 30 frames, resulting in a total of 310 video sequences. To enhance the robustness of the model, data augmentation techniques such as horizontal flipping are applied, ultimately expanding the dataset to 620 video sequences. For model training, the sequences are divided into training, validation, and test sets in a 6:2:2 ratio.

5.2 Experimental environment

The experimental environment described in this paper features a GPU model RTX A5000 with 24 GB of VRAM, accompanied by 14 vCPU Intel(R) Xeon(R) Gold 6330 CPU with 2 GHz.

5.3 Evaluation metrics

Loss refers to the value quantified by the loss function that measures the difference between the model's predicted labels on the test set and the true labels on the test set. It is expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(l_{true} - l_{predicte})^2 \tag{14}$$

where, l_{true} represents the actual label, and $l_{predicte}$ represents the label predicted by the model. In the task of fall prediction, various metrics are used to evaluate the model's performance. The AUC (Area Under the Curve) represents the area under the ROC (Receiver Operating Characteristic) curve, with a value range from 0 to 1. The closer the AUC value is to 1, the better the model's performance. Additionally, the F1 score takes into account both the model's precision (P) and recall (R), and is calculated as:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{15}$$

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \end{cases} \tag{16}$$

where, TN (True Negatives) refers to the number of instances correctly predicted as the negative class (non-fall), while TP (True Positives) refers to the number of instances correctly predicted as the positive class (fall). FN (False Negatives) denotes the number of instances where the positive class is incorrectly predicted as the negative class, and FP (False Positives) refers to the number of instances where the negative class is incorrectly predicted as the positive class.

5.4 Experiments results

5.4.1 Comparing experiments

Human skeletal graphs (Keypoints) and their corresponding dense feature maps (DensePose) are input into the ConvGRU model and the ConvGRU-CBAM model, respectively, for training in a dual-channel manner. During the training process, to prevent overfitting, an early stopping strategy is employed to halt training. The x-axis represents the number of training iterations, while the y-axis indicates the average loss per batch for each iteration. The loss curve initially decreases rapidly, then the rate of decrease slows down and gradually levels off, indicating that the model training is complete. Subsequently, the trained models are tested on a dual-channel test set to evaluate their performance, with the experimental results shown in **Table 3**.

Table 3. Experimental results on a dual-channel test set

Model	MSE	AUC	F1 score
ConvGRU	0.0030	0.9286	0.9181
ConvGRU-CBAM	**0.0041**	**0.9534**	**0.9286**

From the **Table 3**, it's clear that the ConvGRU-CBAM model shows a significant improvement over the ConvGRU model on the dual-channel test set, with a 2.67% increase in AUC and a 0.6% increase in F1 score, among other enhancements. This indicates that the integration of CBAM into the ConvGRU model effectively enhances the model's performance in fall detection tasks.

To validate that the dual-channel input method proposed in this paper can learn more fall features and predict falls more accurately, this experiment also involves training the ConvGRU model and the ConvGRU-CBAM model using human skeletal keypoints or Densepose images as inputs in a single-channel manner.

After the training is deemed complete, as indicated by the two loss curve graphs showing the decrease and stabilization of loss values, the trained models are then tested on a single-channel test set to evaluate their performance. The experimental results are displayed in **Table 4**.

Tabel 4. Experimental results on a single-channel test set

Model	Input	MSE	AUC	F1 score
ConvGRU	Keypoints	0.0099	0.9189	0.9118
ConvGRU	DensePose	0.0089	0.9271	0.9167
ConvGRU-CBAM	Keypoints	0.0048	0.9316	0.9217
ConvGRU-CBAM	DensePose	0.0045	0.9367	0.9242

In **Table 4**, it's noted that under the single-channel dataset, the ConvGRU-CBAM model shows various degrees of improvement over the ConvGRU model. Additionally, comparing **Tables 3** and **4**, the dual-channel ConvGRU-CBAM model outperforms the single-channel ConvGRU-CBAM model with a 2.34% higher AUC and a 0.75% higher F1 score. This further validates that the performance of dual-channel input models is superior to that of single-channel keypoints input models.

To verify whether the dual-channel ConvGRU-CBAM model can predict falls before they actually occur, a fall sequence of 55 frames was used as the input for model prediction. The model outputs the fall probability for each frame and visualizes these probabilities, as illustrated in **Fig. 7**. The original images corresponding to frames 38 to 55 are shown in **Fig. 8**. This approach demonstrates the model's potential in recognizing the precursors to a fall, highlighting its utility in real-world applications where early detection can significantly impact response times and outcomes.

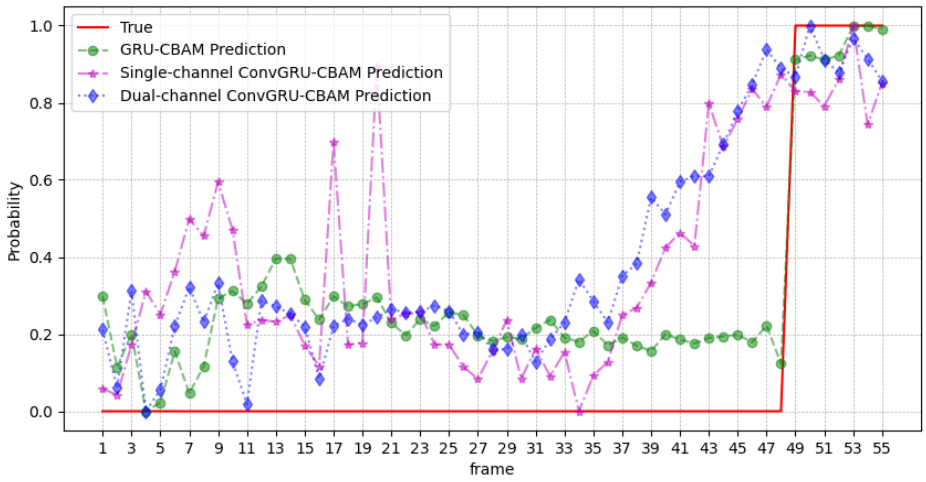


Fig. 7. Fall predictation probability graph from ConvGRU-CBAM dual-channel model

Using keypoint coordinates from the sequence in **Fig. 8** as input for the GRU-CBAM model, and visualizing the fall probability for each frame as shown in **Fig. 7**, a comparison between the red solid line and the green dashed line reveals that the GRU-CBAM model only increases the fall probability and abruptly exceeds the threshold after the fall event starts (around frame 48), indicating it cannot effectively predict falls before they happen. Therefore, it suggests that using dual-channel feature maps as inputs is more suitable for the tasks of this paper, and ConvGRU exhibits stronger learning capabilities compared to GRU.

In **Fig. 7**, the x-axis represents frames 1 to 55, while the y-axis indicates the probability of falling. The red solid line labeled "True" represents the actual probability of falling, the green dashed line labeled "GRU-CBAM Prediction" shows that the fall probability predicted by the single-channel GRU with CBAM. The purple dashed line labeled "Single-channel ConvGRU-CBAM Prediction" illustrates the fall probability predicted by the single-channel ConvGRU-CBAM model, and the blue dashed line labeled "Dual-channel ConvGRU-CBAM Prediction" depicts the fall probability predicted by the dual-channel ConvGRU-CBAM model. In this paper, a probability threshold of 0.5 is set. values above this threshold are interpreted as the occurrence of a fall event, while values below it are considered as no fall event.

Comparing the blue dashed line with the red solid line in **Fig. 7**, it's evident that starting from frame 39, the fall probability predicted by the ConvGRU-CBAM model exceeds the threshold and continues to increase. Thus, it's analyzed that the ConvGRU-CBAM model is capable of predicting fall events before they actually occur, allowing for the implementation of early warnings and protective measures.

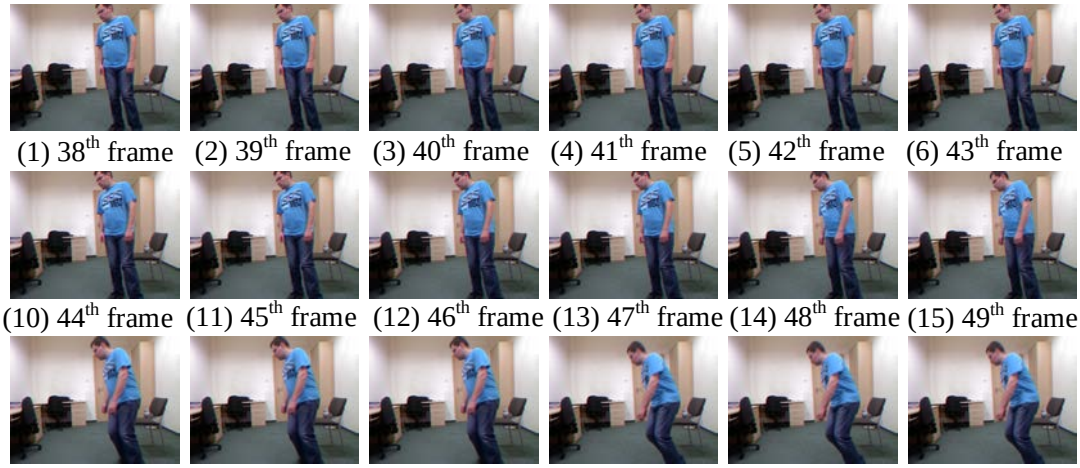


Fig. 8. The original images corresponds to frames 38 to 55

Using the sequence from **Fig. 8** as input for the single-channel ConvGRU-CBAM model, and visualizing the fall probability of each frame as shown in **Fig. 7**, an observation can be made. Comparing the red solid line with the purple dashed line in **Fig. 7**, the single-channel ConvGRU-CBAM model can predict a fall event happening by showing probabilities greater than the threshold starting a few frames before the fall (around frame 43), with the probability of falling increasing over time. However, compared to the purple dashed line (single-channel) and the blue dashed line (dual-channel), fluctuations occur at frames like 9, 17, and 20 in the single-channel model, indicating occasional variability. Together with the results in **Table 4**, it's analyzed that the performance of the single-channel ConvGRU-CBAM model is slightly lower.

5.4.2 Ablation experiments

The experiments conducted so far, utilizing feature maps as model inputs, have demonstrated that dual-channel input models outperform single-channel input models. The objectives of these experiments were to:

(1) Verify that models with ConvGRU as the backbone have stronger learning capabilities than those with GRU as the backbone in the tasks of this paper.
(2) Confirm that in fall prediction tasks, models using feature maps as inputs perform better than models using keypoint coordinates as inputs.
(3) Validate that for the tasks of this paper, using RNNs as the model is more suitable than other deep learning networks.

Models with LSTM, GRU, ConvLSTM and ConvGRU backbones were trained for fall detection using human keypoint coordinate data. Additionally, machine learning models, such as SVM with Radial Basis Function (RBF) kernel and polynomial kernel (Poly), random forest (RF) and k-nearest neighbor (KNN) were used for comparative test. This round of experiments also considered the fusion of channel and spatial features in fall sequences, incorporating CBAM, Spatial Attention Module (SAM), and Channel Attention Module (CAM) for training, and validating on the test set, with results shown in **Table 5**.

The results from **Table 5** indicate that RNNs such as LSTM and GRU as the model backbone perform better in fall detection compared to machine learning based models. Furthermore, when ConvGRU is used as the backbone, integrating both channel and spatial

attention in the CBAM module performs better than using only the CAM or SAM alone.

Table 5. Results of different Backbone models combined with three attention modules

Backbone	Attention	Precision	Recall	F1 score	AUC
SVM (RBF)	-	0.8992	0.6946	0.7837	0.8332
SVM (POLY)	-	0.8540	0.7005	0.7697	0.8287
KNN	-	0.9172	0.7964	0.8525	0.8852
RF	-	0.8766	0.8083	0.8411	0.8836
LSTM	CAM	0.9216	0.8510	0.8848	0.9034
	SAM	0.9500	0.8341	0.8993	0.9093
	CBAM	0.9434	0.8411	0.9061	0.9130
GRU	CAM	0.8952	0.8582	0.8748	0.8939
	SAM	0.8986	0.8592	0.8796	0.8941
	CBAM	0.9086	0.8717	0.8883	0.9179
ConvLSTM	CAM	0.9479	0.8294	0.8875	0.9246
	SAM	0.9416	0.8441	0.8902	0.9304
	CBAM	0.9472	0.8471	0.8944	0.9452
ConvGRU	CAM	0.9514	0.8600	0.9222	0.9355
	SAM	0.9494	0.8501	0.9152	0.9421
	CBAM	**0.9504**	**0.8609**	**0.9286**	**0.9534**

To objectively understand the focus areas of the GRU-CBAM model when analyzing fall detection sequences, this experiment employs the technique of generating attention maps. These maps create an importance matrix of the same size as the keypoint matrix and are visualized as shown in **Fig. 9**. Similarly, the attention maps for daily living sequences (Adl) are visualized, as shown in **Fig. 10**. This visualization technique helps in identifying which regions or keypoints the model pays more attention to during the fall detection process, thereby providing insights into the model's decision-making process.

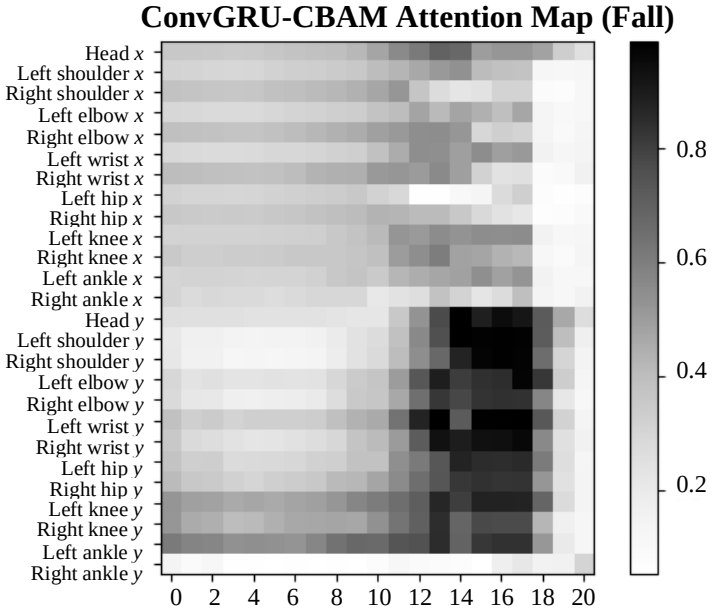


Fig. 9. Attention maps of fall sequences on the ConvGRU-CBAM model

Observing **Fig. 9**, it's evident that the vertical coordinates of keypoints are deemed more important than their horizontal counterparts, indicating that the model focuses more on the vertical changes of the human body. Moreover, for the fall sequence, the importance is greater between frames 10 and 18 than before frame 10, suggesting that the model pays more attention to the frames immediately before and after the fall event occurs.

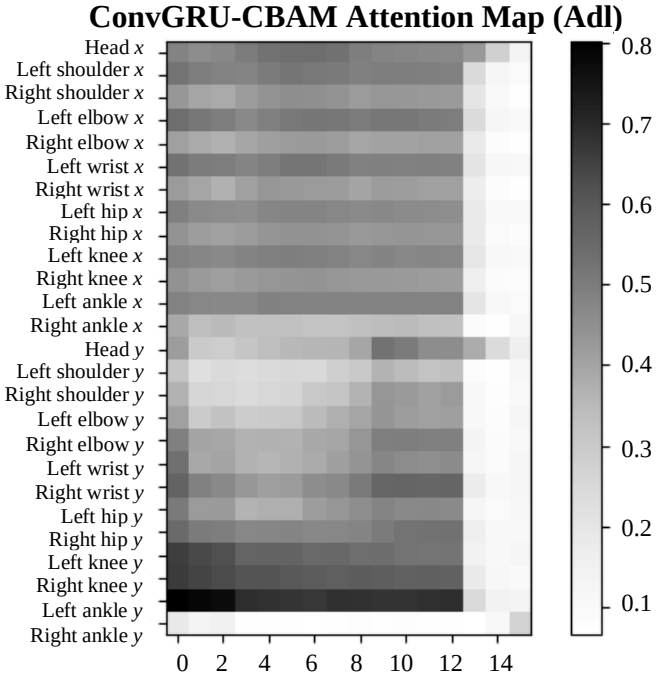


Fig. 10. Attention maps of Adl sequences on the ConvGRU-CBAM model

As shown in **Fig. 10**, since no abnormal behaviors like falls occur in the Adl events, the importance levels of various features are relatively uniform, and the model does not focus on specific frames as it does in **Fig. 9**.

The analysis of attention maps and feature visualization objectively demonstrates the model's learning degree regarding video sequences and highlights the relationship between frames. This insight helps in understanding how the model differentiates between normal activities and fall events by focusing on key moments and spatial changes indicative of a fall.

6. Conclusions

This paper presents a dual-channel algorithmic model for fall detection based on human skeletal keypoint feature maps and dense feature maps. By utilizing human keypoint data extracted via the ViTPose++ algorithm and combining it with dense feature maps generated by the DensePose algorithm within a ConvGRU network model, a novel method for fall detection is formed. This method not only considers information across three dimensions—time (the temporal sequence and relative displacement tracking relationship of ConvGRU), space (the relative position relationship of human keypoints), and channel (the relationship between input feature maps)—but also effectively integrates this information through the CBAM module.

Experimental results show that the dual-channel ConvGRU-CBAM model outperforms the single-channel ConvGRU-CBAM model in both AUC and F1 scores, proving that dual-channel inputs can provide richer feature information, thus enhancing the model's detection performance. Moreover, by integrating time and channel attention in the CBAM, the model can more accurately capture precursor signals of fall events, achieving effective prediction before the occurrence of falls.

Acknowledgements

We thank the Research Fund of Jianghan University (Grant No. 2022SXZX29) and the Hubei Provincial Health and Family Planning Scientific Research Project (Grant No. WJ2023F039) for funding this project.

References

[1] Md. M. Islam et al., "Deep Learning Based Systems Developed for Fall Detection: A Review," *IEEE Access*, vol.8, pp.166117-166137, 2020. Article(CrossRefLink)

[2] Z. Yi, L. Feng, Z. Li, and L. Shouyin, "Human Posture Detection Method Based on Long Short Term Memory Network," *Journal of Computer Applications*, vol.38, no.6, pp.1568-1574, 2018. Article(CrossRefLink)

[3] K. Clarke, T. Ariyarathna, and S. Kumari, "Concept-to-implementation of New Threshold-based Fall Detection Sensor," in *Proc. of TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, pp.597-602, Auckland, New Zealand, Dec. 2021. Article(CrossRefLink)

[4] H. Yhdego, J. Li, C. Paolini, and M. Audette, "Wearable Sensor Gait Analysis of Fall Detection using Attention Network," in *Proc. of 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.3137-3141, Houston, TX, USA, Dec. 2021. Article(CrossRefLink)

[5] S. Mekruksavanich, P. Jantawong, and A. Jitpattanakul, "Pre-Impact Fall Detection Based on Wearable Inertial Sensors using Hybrid Deep Residual Neural Network," in *Proc. of 2022 6th International Conference on Information Technology (InCIT)*, pp.450-453, Nonthaburi, Thailand, Nov. 2022. Article(CrossRefLink)

[6] B. Wang and Y. Guo, "Soft Fall Detection Using Frequency Modulated Continuous Wave Radar And Regional Power Burst Curve," in *Proc. of 2022 Asia-Pacific Microwave Conference (APMC)*, pp.240-242, Yokohama, Japan, Nov. 2022. Article(CrossRefLink)

[7] N. Bharathiraja, R. B. Indhuja, P.R. A. Krishnan, S. Anandhan, and S. Hariprasad, "Real-Time Fall Detection using ESP32 and AMG8833 Thermal Sensor: A Non-Wearable Approach for Enhanced Safety," in *Proc. of 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pp.1732-1736, Trichy, India, Aug. 2023. Article(CrossRefLink)

[8] J.-C. Hou et al., "Cooperative Fall Detection with Multiple Cameras," in *Proc. of 2022 IEEE International Conference on Consumer Electronics - Taiwan*, pp.543-544, Taipei, Taiwan, Jul. 2022. Article(CrossRefLink)

[9] J. Li, Q. Zhao, T. Yang, and C. Fan, "An Algorithm of Fall Detection Based on Vision," in *Proc. of 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*, pp.133-136, Changsha, China, Jun. 2021. Article(CrossRefLink)

[10] Y. Xi, P. Chen, and Z. Fu, "Research on Fall Detection Method of Empty-nesters Based on Computer Vision," in *Proc. of 2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, pp.232-235, Tainan, Taiwan, May 2022. Article(CrossRefLink)

[11] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose++: Vision Transformer for Generic Body Pose Estimation," *arXiv:2212.04246*, Jul. 12, 2023. Accessed: Oct. 12, 2023. Article(CrossRefLink)

[12] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol.117, no.3, pp.489-501, Dec. 2014. Article(CrossRefLink)

[13] R. Rakhimov et al., "Making DensePose fast and light," in *Proc. of 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1868-1876, Waikoloa, HI, USA, Jan. 2021. Article(CrossRefLink)

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. of 15th European Conference on Computer Vision – ECCV 2018*, LNIP, vol.11211, pp.3-19, 2018. Article(CrossRefLink)

[15] B.-S. Lin, C.-W. Peng, I.-J. Lee, H.-K. Hsu, and B.-S. Lin, "System Based on Artificial Intelligence Edge Computing for Detecting Bedside Falls and Sleep Posture," *IEEE J. Biomed. Health Inform.*, vol.27, no.7, pp.3549-3558, Jul. 2023. Article(CrossRefLink)
[16] A. Li et al., "An Integrated Sensing and Communication System for Fall Detection and Recognition Using Ultrawideband Signals," *IEEE Internet Things J.*, vol.11, no.1, pp.1509-1521, 2024. Article(CrossRefLink)
[17] B. Shi, A. Tay, W. L. Au, D. M. L. Tan, N. S. Y. Chia, and S.-C. Yen, "Detection of Freezing of Gait Using Convolutional Neural Networks and Data From Lower Limb Motion Sensors," *IEEE Trans. Biomed. Eng.*, vol.69, no.7, pp.2256-2267, Jul. 2022. Article(CrossRefLink)
[18] K. Nishio, T. Kaburagi, Y. Hamada, T. Matsumoto, S. Kumagai, and Y. Kurihara, "Construction of an Aggregated Fall Detection Model Utilizing a Microwave Doppler Sensor," *IEEE Internet Things J.*, vol.9, no.3, pp.2044-2055, Feb. 2022. Article(CrossRefLink)
[19] V. Divya and R. L. Sri, "Docker-Based Intelligent Fall Detection Using Edge-Fog Cloud Infrastructure," *IEEE Internet Things J.*, vol.8, no.10, pp.8133-8144, May 2021. Article(CrossRefLink)
[20] Z. Ou and W. Ye, "Lightweight Deep Learning Model for Radar-Based Fall Detection With Metric Learning," *IEEE Internet Things J.*, vol.10, no.9, pp.8111-8122, May 2023. Article(CrossRefLink)
[21] M. Musci, D. De Martini, N. Blago, T. Facchinetti, and M. Piastra, "Online Fall Detection Using Recurrent Neural Networks on Smart Wearable Devices," *IEEE Trans. Emerg. Topics Comput.*, vol.9, no.3, pp.1276-1289, Jul.-Sep. 2021. Article(CrossRefLink)
[22] C. Li et al., "Disturbance Propagation Model of Pedestrian Fall Behavior in a Pedestrian Crowd and Elimination Mechanism Analysis," *IEEE Trans. Intell. Transport. Syst.*, vol.25, no.2, pp.1519-1529, 2024. Article(CrossRefLink)
[23] B. Barshan and M. Ş. Turan, "A Novel Heuristic Fall-Detection Algorithm Based on Double Thresholding, Fuzzy Logic, and Wearable Motion Sensor Data," *IEEE Internet Things J.*, vol.10, no.20, pp.17797-17812, Oct. 2023. Article(CrossRefLink)
[24] C. Mosquera-Lopez et al., "Automated Detection of Real-World Falls: Modeled From People With Multiple Sclerosis," *IEEE J. Biomed. Health Inform.*, vol.25, no.6, pp.1975-1984, Jun. 2021. Article(CrossRefLink)
[25] Z. Qian et al., "Development of a Real-Time Wearable Fall Detection System in the Context of Internet of Things," *IEEE Internet Things J.*, vol.9, no.21, pp.21999-22007, Nov. 2022. Article(CrossRefLink)
[26] M. Shen, K.-L. Tsui, M. A. Nussbaum, S. Kim, and F. Lure, "An Indoor Fall Monitoring System: Robust, Multistatic Radar Sensing and Explainable, Feature-Resonated Deep Neural Network," *IEEE J. Biomed. Health Inform.*, vol.27, no 4, pp.1891-1902, Apr. 2023. Article(CrossRefLink)
[27] J. Maitre, K. Bouchard, and S. Gaboury, "Fall Detection With UWB Radars and CNN-LSTM Architecture," *IEEE J. Biomed. Health Inform.*, vol.25, no.4, pp.1273-1283, Apr. 2021. Article(CrossRefLink)
[28] Q. Li, J. Liu, R. Gravina, W. Zang, Y. Li, and G. Fortino, "A UWB-Radar-Based Adaptive Method for In-Home Monitoring of Elderly," *IEEE Internet Things J.*, vol.11, no.4, pp.6241-6252, 2024. Article(CrossRefLink)
[29] C. He et al., "A Noncontact Fall Detection Method for Bedside Application With a MEMS Infrared Sensor and a Radar Sensor," *IEEE Internet Things J.*, vol.10, no.14, pp.12577-12589, Jul. 2023. Article(CrossRefLink)
[30] B. Wang, H. Zhang, and Y.-X. Guo, "Radar-Based Soft Fall Detection Using Pattern Contour Vector," *IEEE Internet Things J.*, vol.10, no.3, pp.2519-2527, Feb. 2023. Article(CrossRefLink)
[31] H. Sadreazami, M. Bolic, and S. Rajan, "Contactless Fall Detection Using Time-Frequency Analysis and Convolutional Neural Networks," *IEEE Trans. Ind. Inf.*, vol.17, no.10, pp.6842-6851, Oct. 2021. Article(CrossRefLink)
[32] Y. Yao et al., "Fall Detection System Using Millimeter-Wave Radar Based on Neural Network and Information Fusion," *IEEE Internet Things J.*, vol.9, no.21, pp.21038-21050, Nov. 2022. Article(CrossRefLink)
[33] F. Jin, A. Sengupta, and S. Cao, "mmFall: Fall Detection Using 4-D mmWave Radar and a Hybrid Variational RNN AutoEncoder," *IEEE Trans. Automat. Sci. Eng.*, vol.19, no.2, pp.1245-1257, Apr. 2022. Article(CrossRefLink)
[34] T. Nakamura, M. Bouazizi, K. Yamamoto, and T. Ohtsuki, "Wi-Fi-Based Fall Detection Using Spectrogram Image of Channel State Information," *IEEE Internet Things J.*, vol.9, no.18, pp.17220-17234, Sep. 2022. Article(CrossRefLink)
[35] Y. Hu, F. Zhang, C. Wu, B. Wang, and K. J. R. Liu, "DeFall: Environment-Independent Passive Fall Detection Using WiFi," *IEEE Internet Things J.*, vol.9, no.11, pp.8515-8530, Jun. 2022. Article(CrossRefLink)

[36] S. Chen, W. Yang, Y. Xu, Y. Geng, B. Xin, and L. Huang, "AFall: Wi-Fi-Based Device-Free Fall Detection System Using Spatial Angle of Arrival," *IEEE Trans. on Mobile Comput.*, vol.22, no.8, pp.4471-4484, Aug. 2023. Article(CrossRefLink)
[37] Z. Yang, Y. Zhang, and Q. Zhang, "Rethinking Fall Detection With Wi-Fi," *IEEE Trans. on Mobile Comput.*, vol.22, no.10, pp.6126-6143, Oct. 2023. Article(CrossRefLink)
[38] Y. Wang, S. Yang, F. Li, Y. Wu, and Y. Wang, "FallViewer: A Fine-Grained Indoor Fall Detection System With Ubiquitous Wi-Fi Devices," *IEEE Internet Things J.*, vol.8, no.15, pp.12455-12466, Aug. 2021. Article(CrossRefLink)
[39] S. Zahan, G. M. Hassan, and A. Mian, "SDFA: Structure-Aware Discriminative Feature Aggregation for Efficient Human Fall Detection in Video," *IEEE Trans. Ind. Inf.*, vol.19, no.8, pp.8713-8721, Aug. 2023. Article(CrossRefLink)
[40] J. Liu, R. Tan, G. Han, N. Sun, and S. Kwong, "Privacy-Preserving In-Home Fall Detection Using Visual Shielding Sensing and Private Information-Embedding," *IEEE Trans. Multimedia*, vol.23, pp.3684-3699, 2021. Article(CrossRefLink)
[41] L. Wu et al., "Video-Based Fall Detection Using Human Pose and Constrained Generative Adversarial Network," *IEEE Trans. Circuits Syst. Video Technol.*, vol.34, no.4, pp. 2179-2194, 2024. Article(CrossRefLink)
[42] R. Zhao et al., "Abnormal Behavior Detection Based on Dynamic Pedestrian Centroid Model: Case Study on U-Turn and Fall-Down," *IEEE Trans. Intell. Transport. Syst.*, vol.24, no.8, pp.8066-8078, Aug. 2023. Article(CrossRefLink)
[43] K. Hu et al., "Graph Fusion Network-Based Multimodal Learning for Freezing of Gait Detection," *IEEE Trans. Neural Netw. Learning Syst.*, vol.34, no.3, pp.1588-1600, Mar. 2023. Article(CrossRefLink)
[44] A. Rezaei, M. C. Stevens, A. Argha, A. Mascheroni, A. Puiatti, and N. H. Lovell, "An Unobtrusive Human Activity Recognition System Using Low Resolution Thermal Sensors, Machine and Deep Learning," *IEEE Trans. Biomed. Eng.*, vol.70, no.1, pp.115-124, Jan. 2023. Article(CrossRefLink)
[45] D. Chen, A. B. Wong, and K. Wu, "Fall Detection Based on Fusion of Passive and Active Acoustic Sensing," *IEEE Internet Things J.*, vol.11, no.7, pp.11566 - 11578, 2024. Article(CrossRefLink)
[46] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. of International Conference on Learning Representations (ICLR 2021)*, 2021. Article(CrossRefLink)
[47] M. Lovanshi and V. Tiwari, "Human Pose Estimation: Benchmarking Deep Learning-based Methods," in *Proc. of 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, pp.1-6, Gwalior, India, Dec. 2022. Article(CrossRefLink)
[48] S. Zhao, Z. Bai, L. Meng, G. Han, and E. Duan, "Pose Estimation and Behavior Classification of Jinling White Duck Based on Improved HRNet," *Animals*, vol.13, no.18, Sep. 2023. Article(CrossRefLink)
[49] S. Reid, S. Coleman, D. Kerr, P. Vance, and S. O'Neill, "Keypoint Changes for Fast Human Activity Recognition," *SN Computer Science*, vol.4, Aug. 2023. Article(CrossRefLink)
[50] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "VSA: Learning Varied-Size Window Attention in Vision Transformers," in *Proc. of 17th European Conference on Computer Vision – ECCV 2022*, LNCS, vol.13685, pp.466-483, 2022. Article(CrossRefLink)
[51] N. Cubero, F. M. Castro, J. R. Cózar, N. Guil, and M. J. Marín-Jiménez, "Multimodal Human Pose Feature Fusion for Gait Recognition," in *Proc. of 11th Iberian Conference on Pattern Recognition and Image Analysis*, LNCS, vol.14062, pp.389-401, 2023. Article(CrossRefLink)
[52] S. Kubo, Y. Iwasawa, M. Suzuki, and Y. Matsuo, "UVTON: UV Mapping to Consider the 3D Structure of a Human in Image-Based Virtual Try-On Network," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp.3105-3108, Seoul, Korea (South), Oct. 2019. Article(CrossRefLink)
[53] Y. Lu, M. Zhang, W. Huang, and S. Guan, "Digitalizing the Dress-up Experience: An Exploration of Virtual Try-On for Traditional Chinese Costume," in *Proc. of 2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, pp.495-500, Ningbo, China, Aug. 2023. Article(CrossRefLink)
[54] T. Liu, H. Xu, and X. Zhang, "3D Clothing Transfer in Virtual Fitting Based on UV Mapping," in *Proc. of 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp.1-6, Beijing, China, Oct. 2018. Article(CrossRefLink)
[55] A. Zhuravlev, "Study of a Method for Effective Noise Suppression in Passive Personnel Screening Systems," in *Proc. of 2019 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)*, pp.1-6, Tel-Aviv, Israel, Nov. 2019. Article(CrossRefLink)

[56] C. Orrite, M. A. Varona, E. Estopiñán, and J. R. Beltrán, "Portrait Segmentation by Deep Refinement of Image Matting," in *Proc. of 2019 IEEE International Conference on Image Processing (ICIP)*, pp.1495-1499, Taipei, Taiwan, Sep. 2019. Article(CrossRefLink)

[57] R. Rijhwani, T. Mahajan, J. Chhatlani, A. Bansode, and G. Bhatia, "Early Diagnosis of Melanoma by Augmenting Feature Extraction of Epidermis using Faster Region-Based Convolutional Neural Networks," in *Proc. of 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, pp.142-147, Hyderabad, India, Sep. 2022. Article(CrossRefLink)

[58] C. Cao, C. Tulvan, M. Preda, and T. Zaharia, "Skeleton-based motion estimation for Point Cloud Compression," in *Proc. of 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp.1-6, Tampere, Finland, Sep. 2020. Article(CrossRefLink)

[59] A. Sanakoyeu, V. Khalidov, M. S. McCarthy, A. Vedaldi, and N. Neverova, "Transferring Dense Pose to Proximal Animal Classes," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5232-5241, Seattle, WA, USA, Jun. 2020. Article(CrossRefLink)

[60] T. Das, S. Sutradhar, M. Das, S. Chakraborty, and S. Deb, "Implementation of a WGAN-GP for Human Pose Transfer using a 3-channel pose representation," in *Proc. of 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pp.698-705, Zallaq, Bahrain, Sep. 2021. Article(CrossRefLink)

[61] R. A. Güler and I. Kokkinos, "HoloPose: Holistic 3D Human Reconstruction In-The-Wild," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10876-10886, Long Beach, CA, USA, Jun. 2019. Article(CrossRefLink)

[62] Y. Zhou, J. Deng, and S. Zafeiriou, "Improve Accurate Pose Alignment and Action Localization by Dense Pose Estimation," in *Proc. of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp.480-484, Xi'an, China, May 2018. Article(CrossRefLink)

[63] A. Ianina, N. Sarafianos, Y. Xu, I. Rocco, and T. Tung, "BodyMap: Learning Full-Body Dense Correspondence Map," in *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.13276-13285, New Orleans, LA, USA, Jun. 2022. Article(CrossRefLink)

[64] J. Zhang, A. Yang, C. Miao, X. Li, R. Zhang, and D. N. H. Thanh, "3D Graph Convolutional Feature Selection and Dense Pre-Estimation for Skeleton Action Recognition," *IEEE Access*, vol.12, pp.11733-11742, 2024. Article(CrossRefLink)

[65] X. Gao and D. Chen, "Pose estimation for six-axis industrial robots based on pose distillation," in *Proc. of 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, pp.105-109, Wuhan, China, Apr. 2022. Article(CrossRefLink)

[66] X. Shi et al., "Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model," in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. Article(CrossRefLink)

[67] Y. Zheng, R. Xiao, and Q. He, "Human Keypoint-Guided Fall Detection: An Attention-Integrated GRU Approach," in *Proc. of the 2023 4th International Conference on Machine Learning and Computer Application (ICMLCA 2023)*, pp.221-226, Hangzhou, China. 2024. Article(CrossRefLink)

Yi Zheng received the M.S. and Ph.D. degrees from the College of Physical Science and Technology, Central China Normal University, Wuhan, China, in 2018 and 2022, respectively. He is currently a Assistant Professor with the Intelligent Medical Engineering Research Center, the School of Artificial Intelligence, Jianghan University, China. His main research interests include intelligent medicine and deep learning.

Cunyi Liao received the B.S. degree and the M.S. degree from the College of Physical Science and Technology, Central China Normal University, China, in 2020 and 2023, and study for Ph.D. degree at the College of Physical Science and Technology, Central China Normal University, China. His research interests include wireless communication and deep learning.

Ruifeng Xiao received the B.S. degrees from the College of Information Engineering, Qingdao Institute of Technology, Shandong, China, in 2021. She is pursuing the M.S. degree at the Intelligent Medical Engineering Research Center of the School of Artificial Intelligence at Jianghan University. Her research interests include electronic information engineering.

Qiang He holds the Ph.D. in Computer Science from the University of Louisiana at Lafayette, U.S.. He is the director for the Intelligent Medical Engineering Research Center and professor at the School of Artificial Intelligence, Jianghan University, China. His research interests lie in Big Data Processing, Intelligent Systems, and their applications.