Pattern Recognition Letters 185 (2024) 31–37

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Fall detection algorithm based on global and local feature extraction

Bin Li [a,*], Jiangjiao Li [b], Peng Wang [a]

[a] School of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences), Jinan, 250353, Shandong, China
[b] School of Information Science and Engineering, Northeastern University, Shenyang, 110819, Liaoning, China

ARTICLE INFO

Editor: Francesco Fontanella

Keywords:
Dual-stream network
Convolutional neural network
Regional attention module
Transformer

ABSTRACT

Falls have become one of the main causes of injury and death among the elderly. A high-accuracy fall detection method can effectively detect falls in the elderly, thereby reducing the probability of injury and mortality. This paper proposes a fall detection algorithm based on global and local feature extraction. Specifically, we design a dual-stream network, with one branch composed of a convolutional neural network and a regional attention module for extracting local features from images. The other branch consists of an improved Transformer for extracting global features from images. The local and global features are then fused using a feature fusion module for classification, enabling fall detection. Experimental results show that the proposed approach achieves accuracies of 99.55% and 99.75% when tested with UP-Fall Detection Dataset and Le2i Fall Detection Dataset.

1. Introduction

Fall-related injuries pose a significant public health concern worldwide. It is estimated that 684,000 fatal fall-related incidents occur annually [1], making falls the second leading cause of unintentional injury deaths, surpassed only by road traffic injuries. More than half of the people who die each year due to falls are people over the age of 60. Consequently, timely detection and rescue following falls in the elderly have become a focal point of concern.

In recent years, machine learning and deep learning algorithms have been widely applied in the field of fall detection. In the realm of deep learning, existing fall detection methods can be classified into three categories: the wearable based, the environment based, and computer vision based.

Fall detection systems based on wearable sensors typically utilize accelerometers, gyroscopes, and other sensors to gather data for fall detection based on metrics such as velocity and acceleration. Kerdjidj et al. [2] proposed a fall detection hardware framework implemented on Zedboard FPGA using accelerometer and gyroscope data. Alarifi and Alwadain [3] presented a fall detection system based on a magnetometer, gyroscope, and accelerometer tri-axial setup. Jansi and Amutha [4] proposed a method for fall detection using tri-axial data from an accelerometer and depth maps from a Kinect sensor. However, wearable sensors are prone to being forgotten or uncomfortable for the elderly to wear, and may sometimes be rendered unusable due to low battery power. As a result, researchers have proposed environment-based fall detection systems. Environment-based fall detection systems typically employ non-visual sensors such as radar, infrared, and ultrasound to gather data for fall detection. Wang et al. [5] introduced a fall detection method based on millimeter-wave frequency modulated continuous wave Radar. Chen et al. [6] proposed a three-tier low-complexity human fall detection method based on IR-UWB radar. Yang et al. [7] presented a fall detection system based on an infrared array sensor and multi-dimensional feature fusion. However, environment-based sensors are susceptible to external interference and exhibit higher false detection rates. Consequently, researchers have proposed computer vision-based fall detection systems. Computer vision-based fall detection systems typically employ devices such as smartphones or cameras to capture images or videos for fall detection. Feng et al. [8] proposed an image-based fall detection method using YOLOv3, Convolutional Neural Networks (CNNs) and Long Short-Term Memory network for feature extraction and classification. Saurav et al. [9] introduced a video-based dual-stream fusion neural network for fall detection. Li et al. [10] proposed a video-based expanded spatiotemporal convolutional autoencoder for fall detection.

The aforementioned fall detection methods all utilize sensors to gather data and employ deep learning techniques for feature extraction and classification. However, these methods exhibit poor specificity and limited ability to extract global features. Therefore, this paper proposes a fall detection algorithm based on global and local feature extraction. We design a two-stream network, one branch is used to extract local features of images, which mainly consists of a convolutional neural network and a regional attention module. Another branch is used to extract global features of images, which mainly consists of an improved

* Corresponding author.
E-mail addresses: binli@qlu.edu.cn (B. Li), 2390100@stu.neu.edu.cn (J. Li), 202011110042@stu.qlu.edu.cn (P. Wang).

https://doi.org/10.1016/j.patrec.2024.07.003
Received 17 March 2024; Received in revised form 7 June 2024; Accepted 2 July 2024
Available online 8 July 2024

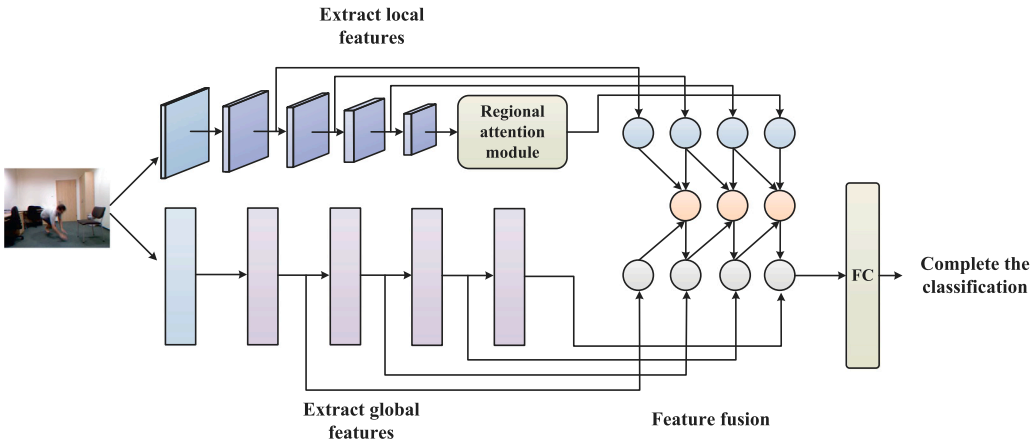


Fig. 1. Flowchart of the proposed fall detection algorithm. Firstly, the input image is processed by the branch consisting of a convolutional neural network and a regional attention module to extract local features. Next, the branch based on the improved Transformer is employed to extract global features from the input image. Finally, the local and global features are fused, and the classification task is performed.

Transformer. The feature fusion module is used to fuse the local features and global features and use them in the fully connected layer to complete the classification of falls and normal. The main contributions of this paper are as follows:

1. A dual-stream fall detection algorithm is proposed, which extracts both global and local features. One branch primarily consists of a convolutional neural network and our design regional attention module, while the other branch mainly consists of our improved Transformer.

2. A feature fusion module is designed to combine local and global features. The local features have more detailed information, and the global features have more semantic information. The fused features retain better detail information and semantic information at the same time, improving the accuracy of the classification results.

3. Extensive computational experiments are conducted on the UP-Fall and Le2i datasets. At the same time, the self-built dataset is used to verify the proposed fall detection algorithm. The experimental results demonstrate that our network exhibits good specificity, sensitivity, and accuracy.

The remaining sections of this paper are organized as follows. Section 2 provides an overview of related work. Section 3 presents the algorithm proposed in this paper. Section 4 describes and discusses the experimental results obtained from the algorithm on publicly available datasets. Section 5 concludes the paper, providing insights into future research directions for fall detection.

2. Related work

Convolutional neural network is a commonly used feature extraction method in fall detection. This section introduces a fall detection algorithm based on convolutional neural network. At the same time, the application of Transformer in computer vision is introduced.

2.1. Convolution-based fall detection algorithm

CNNs are a type of feedforward neural network that consist of multiple convolutional and pooling layers. CNNs have demonstrated outstanding performance in image processing tasks. Currently, many fall detection systems rely on CNNs to extract image features for classification.

Gao et al. [11] proposed an image-based fall detection method using Openpose and MobileNetV2 for feature extraction and classification. Li et al. [12] utilized a convolution-based adaptive keypoint attention module and an improved residual long short-term attention network to extract spatiotemporal features, achieving binary classification for falls and normal activities. Xu et al. [13] combined a thresholding approach with a convolutional neural network in an algorithm applied for real-time fall detection on wearable devices.

2.2. Transformer in computer vision

The Transformer was introduced by Vaswani et al. [14] in 2017 and was initially applied primarily in natural language processing. The Transformer is the first network that relies entirely on self-attention mechanisms, abandoning traditional convolutional neural networks and recurrent neural networks. As a result, the Transformer is capable of efficient parallel computation and extracting contextual information and global features effectively.

The Transformer was formally applied to the field of computer vision in 2020 and has achieved promising results. The Vision Transformer (ViT) proposed by Dosovitskiy et al. [15] is a network that uses Transformers for image classification. They employ Transformer encoder modules for feature extraction from image patches, and multilayer perceptrons are utilized for classification. However, ViT operates on the flattened feature map of the entire image, resulting in a large number of network parameters and high computational complexity. To address this issue, Liu et al. [16] propose the Shifted Windows Transformer, which introduces downsampling operations and shifted windows on top of ViT, enabling encoding operations within the shifted windows for information interaction and reducing the computational complexity of the network. Similarly, based on the aforementioned limitations of ViT, Bai et al. [17] propose the compression methods for the Multi-Head Self-Attention module to reduce the FLOPs of ViT. They also propose a pruning method named Multi-Dimension Compression of Feed-Forward Network in Vision Transformers to reduce the computational complexity of ViT.

Based on the application of the above-mentioned convolutional neural network and Transformer, this paper proposes a fall detection algorithm based on global and local feature extraction, which overcomes the poor ability of the convolutional neural network to extract global features and the poor ability of the Transformer to extract local features for fall detection.

3. The proposed method

This section presents the proposed fall detection algorithm based on the extraction of global and local features. It consists of three main components. The first part showcases a branch composed of a convolutional neural network and a regional attention module, which is responsible for extracting the local features of the image. The second part demonstrates our improved Transformer, which is utilized for capturing the global features of the image. The third part introduces the feature fusion module, responsible for integrating the local and global features of the image. The fused features are then used for fall detection classification. The workflow of the aforementioned fall detection algorithm is illustrated in Fig. 1.

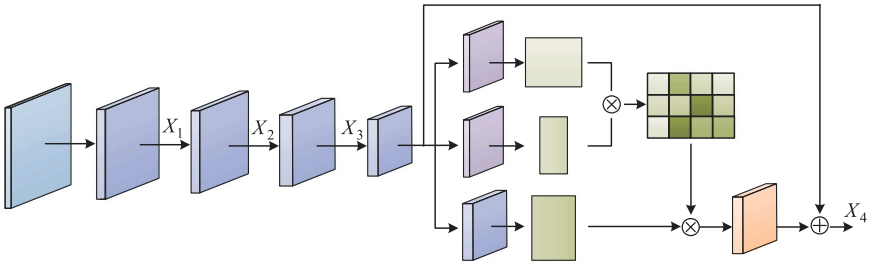


Fig. 2. Branches based on Convolutional Neural Networks and Regional Attention Modules.

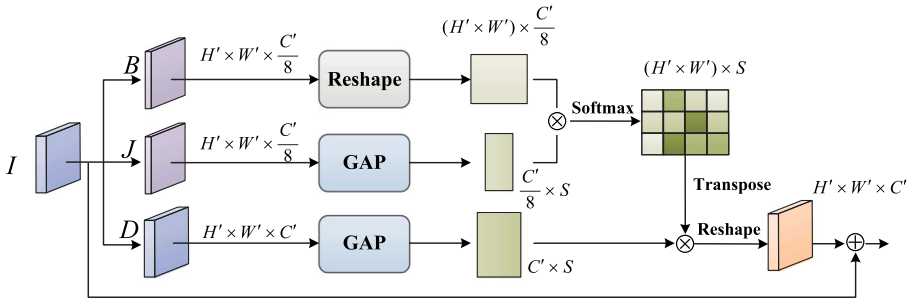


Fig. 3. Regional attention module structure.

3.1. Branches for extracting local features

Before inputting the dual-stream network, we use the following method to perform batch normalization processing on the data T of the input $H \times W \times C$ to realize the normalization processing of the samples and accelerate the network convergence at the same time. Eq. (1) [18] represents the normalized samples:

$$T' = \frac{T - \mu}{\sqrt{\sigma^2 + \varepsilon}} * \gamma + \beta \tag{1}$$

where, T represents the sample, μ represents the mean of the normalized data of the layer, σ represents the variance of the normalized data of the layer, ε is a small constant that prevents the denominator from being zero, γ and β are the learning parameters of the model, taking 1 and 0 respectively.

The advantage of convolutional neural networks lies in their ability to extract image features using convolutional kernels and reduce redundant information through pooling operations, thereby reducing the dimensionality of feature maps. Based on the above, it can be concluded that convolutional neural networks are more suitable for extracting smaller and local features. Therefore, we have designed a branch based on convolution and regional attention module to extract local features from the data.

We input the dataset T' into the branch based on convolution and regional attention module, and the overall structure of the branch is shown in Fig. 2. The dataset T' undergoes a 7×7 convolution operation followed by max pooling, which serves as a preprocessing step for the dataset, as shown in Eq. (2):

$$x = W_1 f_{7\times7}\left(T'\right) \tag{2}$$

where, $f_{7\times7}$ represents the 7×7 convolution operation, W_1 represents the max pooling operation. Then, x undergoes the residual block structure in the ResNet network [19], where a residual block consists of two 3×3 convolutions. A residual connection connects the two convolutional layers, effectively addressing the issue of gradient vanishing in the network, as shown in Eq. (3):

$$X_1 = f_{3\times3}(x) + x \tag{3}$$

where, $f_{3\times3}$ represents two 3×3 convolutional operations. Each time the residual structure is passed, an output is obtained, thus obtaining X_2 and X_3. After passing through four consecutive residual blocks, a feature map I of dimension $H' \times W' \times C'$ is obtained.

The attention mechanism can give greater weight to the important parts of the feature map, and less weight to the unimportant parts, so as to put more attention on the more important parts of the map. Therefore, this paper proposes the regional attention modules to better extract local features. The feature map is input into our designed regional attention module. The feature map I is treated separately as B, J, D, where B, J perform channel dimension reduction, resulting in $B, J \in R^{H' \times W' \times \frac{C'}{8}}$ and reducing computational complexity. J, D undergoe global average pooling (GAP), yielding $J' \in R^{\frac{C'}{8} \times S}$ and $D' \in R^{C' \times S}$. After reshaping, the dimension of B becomes $(H' \times W') \times \frac{C'}{8}$. The element-wise multiplication between B and J', followed by a softmax operation, yields an attention matrix E with dimension $(H' \times W') \times S$, as shown in Eq. (4):

$$E = \mathrm{softmax}(\mathrm{Re}(B) \otimes GAP(C)) \tag{4}$$

where, Re represents the reshaping operation, GAP represents the global average pooling operation, $\otimes$ represents the matrix multiplication operation, and softmax represents the activation function. E is then element-wise multiplied with D, resulting in F with dimension $C' \times (H' \times W')$. F is reshaped and added to I, yielding the final output X_4, as shown in Eq. (5):

$$X_4 = I \oplus \mathrm{Re}\left(E \otimes \left(W_2 D\right)\right) \tag{5}$$

where, $\oplus$ represents the matrix addition operation. The overall process of the regional attention module is illustrated in Fig. 3. Through the aforementioned convolutional and regional attention modules, the extraction of local features from the dataset is accomplished.

3.2. Branch for extracting global features

The advantage of the Transformer is that there is a self-attention mechanism, which can effectively obtain global information, and multiple heads can map it to multiple spaces, making the model's expressive ability stronger. It breaks through the limitation that RNN needs to be executed in a loop and cannot be calculated in parallel. Therefore, Transformer is more suitable for extracting larger features and

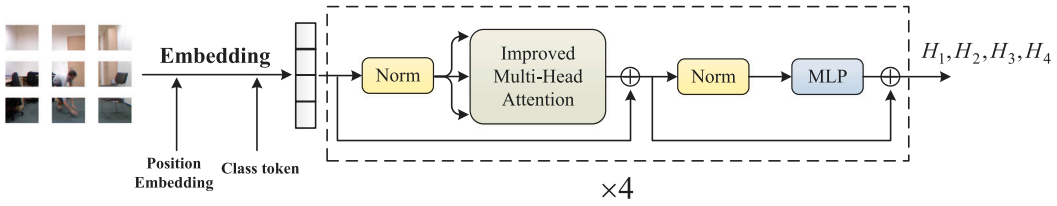


Fig. 4. Branch based on improved Transformer.

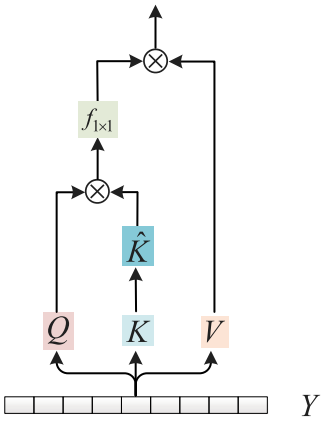


Fig. 5. Improved multi-head self-attention module structure.

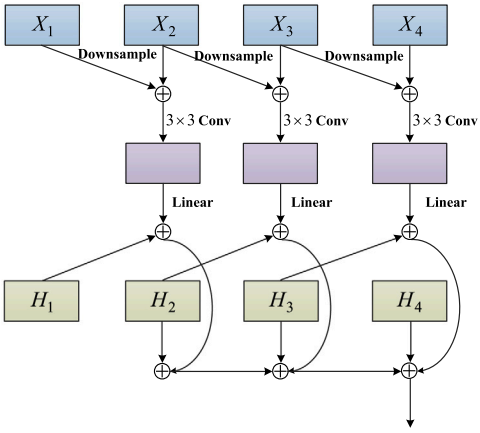


Fig. 6. Feature fusion module structure.

full features. Therefore, we design a branch based on the improved Transformer for extracting global features in the data.

We partition the normalized images into blocks and input them into the branch of the improved Transformer, as shown in Fig. 4. We divide the images into blocks of size $P\times P$, resulting in $N = \frac{HW}{P^2}$ image blocks, each with dimensions of $P^2 \times C$. Then, we perform embedding on the image blocks, which means applying a linear transformation to each $P^2\times C$ image block to compress its dimensions to $\bar{D}$. This yields a tensor of $N \times \bar{D}$. Additionally, we use the positional encoding method from Vision Transformer [15] to generate positional encodings of dimension $N \times \bar{D}$. These positional encodings are added element-wise to a tensor of the same dimensions, resulting in the network input of dimension $N \times \bar{D}$, denoted as Y. The tensor Y undergoes layer normalization and becomes Y'.

Then, Y' is processed by the improved multi-head self-attention module, as illustrated in Fig. 5. The original self-attention module only uses the current Query-Key pair to calculate the attention matrix, ignoring the contextual information between different keys. To address this, we introduce feature interaction and convolutional operations to the original self-attention module. First, we perform feature interaction K, which means that each value in the matrix undergoes a feature fusion operation with its adjacent values, resulting in $\hat{K} \in N \times \bar{D}$. Then, we multiply Q and K' together. Finally, the attention matrix is multiplied by V after convolution $f_{1\times1}$ (two 1×1-dimensional convolutions and a residual structure), resulting in the output of a self-attention head. The outputs of the multi-head self-attention are represented as G. G undergoes further processing by a multi-layer perceptron module and produces the output of the improved Transformer encoding module, denoted as H_4, as shown in Eq. (6):

$$H_4 = \left(Y \oplus m\left(Y'\right)\right) \oplus \left(f\left(l\left(Y \oplus m\left(Y'\right)\right)\right)\right) \tag{6}$$

where, m represents the multi-head self-attention operation, l represents the layer normalization operation, and f represents the multi-layer perceptron operation. By stacking four improved Transformer encoding modules, the extraction of global features from the image is completed.

3.3. Feature fusion

To integrate the local features extracted by the two branches and the global features, we have designed a feature fusion module to compensate for the limitations of convolutional neural networks in extracting global features and the limitations of Transformers in extracting local features.

Input the extracted local features and global features into the feature fusion module. The structure of the feature fusion module is shown in Fig. 6. Local features are fused with global features through downsampling, convolution, and linear transformation operations. The downsampling operation makes the dimensions of high-level features and low-level features consistent, thereby performing feature fusion of different levels of local features. The 3×3 convolution operation can transform the summed features so that the fusion of high-level features and low-level features is more in-depth. The linear operation makes the dimensions of the corresponding layer of the local features consistent with the global features, thereby performing local and global feature fusion. The output of the feature fusion module is applied to the fully connected layer (FC). To prevent the model from overfitting and improve the generalization ability of the model, the dropout layers are added after the fully connected layer. The dropout rate is 0.3. The fully connected layer completes the two classifications of falling and normal.

4. Experimental results

In terms of hardware, the experiments are conducted on a machine equipped with an Intel Xeon E5-2698 v4 processor, 256 GB of memory, and an NVIDIA Tesla V100. In terms of software, the operating system used was Ubuntu 18.04, the programming language was Python 3.7, and the deep learning library employed was PyTorch 1.7.0. We train the algorithm for 100 epochs on the UP-Fall and Le2i datasets, using the cross-entropy loss function, the batch is set to 64, the learning rate is initially 0.001, and every 20 epochs becomes 0.5 of the original learning rate.

Table 1
Ablation experiment to extract local feature branch.

Structure	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Fully Connected Layers	93.13	90.52	88.27	91.31
Fully Connected Layers + CNN	98.32	97.75	96.56	98.47
Fully Connected Layers + CNN + Regional attention module	**99.14**	**98.52**	**96.97**	**98.61**

Table 2
Ablation experiment to extract global feature branch.

Structure	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Fully Connected Layers	93.63	90.46	89.82	90.27
Fully Connected Layers + Transformer	98.89	98.38	96.94	98.22
Fully Connected Layers + Improved Transformer	**99.32**	**98.63**	**97.09**	**98.76**

Table 3
Performance comparison between dual-stream network and single-stream network.

Structure	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Branch for extracting local features	99.14	98.52	96.97	98.61
Branch for extracting global features	99.32	98.63	97.09	98.76
Dual-stream network	**99.55**	**99.24**	**97.12**	**99.08**

Table 4
Results of different network structures on the UP-Fall dataset.

Method	Feature extraction	Sensitivity (%)	Specificity (%)	Accuracy (%)
nez Marcos and Arganda-Carreras [20]	Transformer	–	–	96.67
Ramirez et al. [21]	Skeleton Features	–	99.71	99.50
Yadav et al. [22]	Spatial and temporal features	–	–	96.70
Zhao et al. [23]	Multi-scale temporal features	95.43	99.12	98.85
Proposed method	Global and local	**98.12**	**99.86**	**99.55**

4.1. Dataset description

4.1.1. UP-Fall dataset

The UP-Fall dataset [24] is a large fall detection dataset composed of 11 activities, each with 3 trials, and recorded using 17 young adults without impairments. Among them, the 11 activities include falling forward using hands, falling forward using knees, falling backwards, falling sideward, falling sitting in an empty chair, walking, standing, sitting, sicking up an object, jumping and laying.

4.1.2. Le2i dataset

The Le2i dataset [25] is released by the Le2i laboratory at the University of Burgundy, France. It includes a total of 82 downloadable videos captured in four main scenes: cafe, home, classroom, and office. The videos are primarily recorded by cameras positioned horizontally to the ground.

4.1.3. Dataset collected by ourselves

In this study, a dataset for fall detection was collected, which was obtained in a laboratory environment. The dataset was intentionally made challenging due to the complex background environment and insufficient lighting in the laboratory. It consists of 30 videos, with each video containing approximately 150–200 frames of images.

4.2. Evaluation metrics

When detecting a video sequence, four possible cases are corresponding to four valid parameters:

True positive (TP): the label is a fall, and is correctly detected as a fall.

False positive (FP): the label is normal, and is incorrectly detected as a fall.

True negative (TN): the label is normal, and is correctly detected as normal.

False negative (FN): the label is a fall, and is incorrectly detected as normal.

Accuracy, precision, recall, specificity, sensitivity, and F-score are commonly used as performance metrics, so the experiments in this paper also use these metrics to evaluate the model.

4.3. Ablation experiment

To validate the effectiveness of the designed regional attention module, improved self-attention mechanism module, and two-stream network, we conducted ablation experiments using the UP-Fall dataset. First, we performed ablation experiments on the branch that extracts local features from images based on CNN and the regional attention module. As shown in Table 1, it can be observed that the designed regional attention module provided a certain improvement to the network's results. Next, we conducted ablation experiments on the branch that extracts global features from images using the improved Transformer, as shown in Table 2. As depicted in Table 2, the improved Transformer brought about a certain improvement in the network's results. Lastly, we validated the effectiveness of the two-stream network. As shown in Table 3, it can be observed that compared to the network with only a single stream, the two-stream network demonstrated better performance.

4.4. Model evaluation on UP-Fall dataset

This section presents the experimental results of the fall detection model on the UP-Fall dataset. For the UP-Fall dataset, the training set and test set were divided in a 7:3 ratio. Figs. 7(a) and 7(b) show the accuracy and loss curves of the proposed algorithm, demonstrating good accuracy and fast convergence.

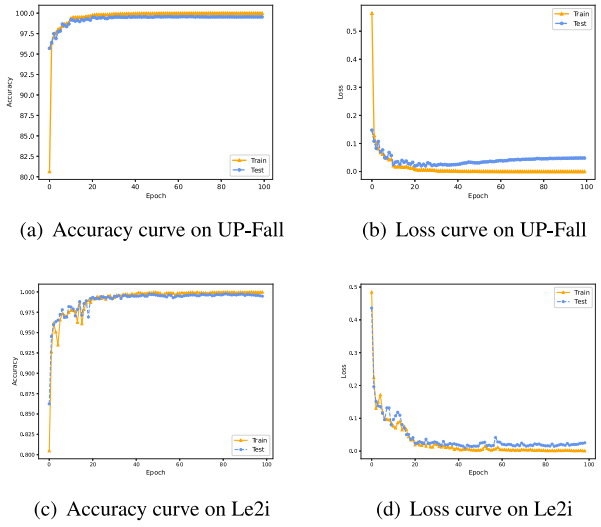
The results were compared with [20–23], and the comparison results are shown in Table 4. From the table, it can be observed that the proposed algorithm achieved better accuracy and specificity. Moreover, compared to [21,22], our method extracted more global features. While [21,22] used CNN to extract local features from the images, our approach used a CNN branch to extract local features and a Transformer branch to extract global features, followed by feature fusion. Therefore, our method captures more comprehensive features.

Table 5
Results of different network structures on the Le2i dataset.

Method	Feature extraction	Sensitivity (%)	Specificity (%)	Accuracy (%)
Beddiar et al. [26]	Angle and distance	–	–	96.20
Nasir et al. [27]	SqueezeNet	95.35	94.34	94.79
Gao et al. [11]	Skeleton Features	–	–	98.60
Zhang et al. [28]	Angle Difference-value	–	–	96.21
Proposed method	Global and local	**99.19**	**99.76**	**99.75**

Table 6
The results of the algorithm in this paper and other algorithms on the dataset collected by ourselves.

Method	Feature extraction	Sensitivity (%)	Specificity (%)	Accuracy (%)
Gao et al. [11]	Keypoints	90.37	89.86	90.98
Beddiar et al. [26]	Deep feature	87.64	88.65	88.36
Galvão et al. [29]	Temporal features	88.48	89.69	88.97
Proposed method	Global and local	**91.73**	**92.36**	**93.41**



(a) Accuracy curve on UP-Fall (b) Loss curve on UP-Fall

(c) Accuracy curve on Le2i (d) Loss curve on Le2i

Fig. 7. Accuracy and loss curves on UP-Fall and Le2i datasets.

4.5. Model evaluation on the Le2i dataset

This section presents the experimental results of the fall detection model on the Le2i dataset. For the experiments on the Le2i dataset, the training set and test set were divided in a 7:3 ratio, with 24,699 samples in the training set and 10,585 samples in the test set. Figs. 7(c) and 7(d) show the accuracy and loss curves of the proposed algorithm.

The results were compared with [11,26–28], and the comparison results are shown in Table 5. From the table, it can be observed that our approach achieved better accuracy, sensitivity, and specificity. At the same time, compared with [26,28], our method is less affected by illumination and viewing angle. By employing CNN, the Regional attention module, and the Transformer, our method processes the images. The Regional attention module focuses more attention on important regions, such as regions containing people, while paying less attention to background information, thereby reducing the impact of the background. The Transformer has an excellent ability to extract global features, enabling good feature extraction and detection performance even in poorly lit images. Therefore, our method is less affected by lighting conditions and viewing angles.

4.6. Model validation

To further test the effectiveness of the algorithm, the proposed network was validated on the collected dataset. The validation set consisted of 4632 images with a size of 320 × 240. Without training on this dataset, the algorithm achieved an accuracy of 93.41%. Table 6 presents the results of reproducing several other fall detection methods proposed by different researchers, and compares them with the results of Gao et al. [11]; Beddiar et al. [26]; Galvão et al. [29]. Our network achieved better accuracy, sensitivity, and specificity.

5. Conclusion

This paper proposes a two-stream network to extract local and global features for fall detection. Among them, the branch for extracting local features consists of a convolutional neural network and a regional attention module. The regional attention module can give greater weight to more important parts of the image to improve the ability of this branch to extract local features. The branch that extracts global features consists of an improved Transformer. We introduce feature interaction and convolution operations in the multi-head self-attention mechanism module to improve the ability of this branch to extract global features. The experimental section compares with current state-of-the-art algorithms, thereby demonstrating the effectiveness of the dual-stream network. However, there are areas in which the algorithm can be further improved. For instance, the network struggles to effectively distinguish between lying down after a fall and lying down in daily life scenarios. Therefore, in future research, our focus will be on differentiating between post-fall lying down and lying down during regular activities.

CRediT authorship contribution statement

Bin Li: Writing – review & editing, Methodology, Conceptualization. **Jiangjiao Li:** Writing – original draft, Validation, Software, Methodology, Data curation. **Peng Wang:** Writing – review & editing, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Colleges and Universities Twenty Terms Foundation of Jinan City, China (2021GXRC100) and the Natural Science Foundation of Shandong Province, China under Grant ZR2020MF097.

References

[1] X. Chen, L. He, K. Shi, J. Yang, X. Du, K. Shi, Y. Fang, Age-stratified modifiable fall risk factors in Chinese community-dwelling older adults, Arch. Gerontol. Geriat. 108 (2023) 104922.

[2] O. Kerdjidj, E. Boutellaa, A. Amira, K. Ghanem, F. Chouireb, A hardware framework for fall detection using inertial sensors and compressed sensing, Microprocess. Microsyst. 91 (2022) 104514.

[3] A. Alarifi, A. Alwadain, Killer heuristic optimized convolution neural network-based fall detection with wearable IoT sensor devices, Measurement 167 (2021) 108258.

[4] R. Jansi, R. Amutha, Detection of fall for the elderly in an indoor environment using a tri-axial accelerometer and kinect depth data, Multidimens. Syst. Signal Process. 31 (4) (2020) 1207–1225.

[5] B. Wang, Z. Zheng, Y.-X. Guo, Millimeter-wave frequency modulated continuous wave radar-based soft fall detection using pattern contour-confined Doppler-time maps, IEEE Sens. J. 22 (10) (2022) 9824–9831.

[6] M. Chen, Z. Yang, J. Lai, P. Chu, J. Lin, A three-stage low-complexity human fall detection method using IR-UWB radar, IEEE Sens. J. 22 (15) (2022) 15154–15168.

[7] Y. Yang, H. Yang, Z. Liu, Y. Yuan, X. Guan, Fall detection system based on infrared array sensor and multi-dimensional feature fusion, Measurement 192 (2022) 110870.

[8] Q. Feng, C. Gao, L. Wang, Y. Zhao, T. Song, Q. Li, Spatio-temporal fall event detection in complex scenes using attention guided LSTM, Pattern Recognit. Lett. 130 (2020) 242–249.

[9] S. Saurav, R. Saini, S. Singh, A dual-stream fused neural network for fall detection in multi-camera and 360 videos, Neural Comput. Appl. 34 (2) (2022) 1455–1482.

[10] S. Li, X. Song, S. Xu, H. Qi, Y. Xue, Dilated spatial–temporal convolutional auto-encoders for human fall detection in surveillance videos, ICT Express 9 (4) (2023) 734–740.

[11] M. Gao, J. Li, D. Zhou, Y. Zhi, M. Zhang, B. Li, Fall detection based on OpenPose and MobileNetV2 network, IET Image Process. 17 (3) (2023) 722–732.

[12] J. Li, M. Gao, B. Li, D. Zhou, Y. Zhi, Y. Zhang, KAMTFENet: a fall detection algorithm based on keypoint attention module and temporal feature extraction, Int. J. Mach. Learn. Cybern. 14 (5) (2023) 1831–1844.

[13] T. Xu, H. Se, J. Liu, A fusion fall detection algorithm combining threshold-based method and convolutional neural network, Microprocess. Microsyst. 82 (2021) 103828.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv. Neural Inf. Process. Syst., 30, 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale. CoRR abs/2010.11929, CoRR (2020) arXiv:2010.11929.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 10012–10022.

[17] S. Bai, J. Chen, Y. Yang, Y. Liu, Multi-dimension compression of feed-forward network in vision transformers, Pattern Recognit. Lett. 176 (2023) 56–61.

[18] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, vol. 37, 2015, pp. 448–456.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[20] A.N. nez Marcos, I. Arganda-Carreras, Transformer-based fall detection in videos, Eng. Appl. Artif. Intell. 132 (2024) 107937.

[21] H. Ramirez, S.A. Velastin, S. Cuellar, E. Fabregas, G. Farias, BERT for activity recognition using sequences of skeleton features and data augmentation with GAN, Sensors 23 (3) (2023) 1400.

[22] S.K. Yadav, A. Luthra, K. Tiwari, H.M. Pandey, S.A. Akbar, ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling, Knowl.-Based Syst. 239 (2022) 107948.

[23] Z. Zhao, L. Zhang, H. Shang, A lightweight subgraph-based deep learning approach for fall recognition, Sensors 22 (15) (2022) 5482.

[24] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez Martínez, C. Peñafort-Asturiano, UP-fall detection dataset: A multimodal approach, Sensors 19 (9) (2019) 1988.

[25] I. Charfi, J. Miteran, J. Dubois, M. Atri, R. Tourki, Definition and performance evaluation of a robust SVM based fall detection solution, in: 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012, pp. 218–224.

[26] D.R. Beddiar, M. Oussalah, B. Nini, Fall detection using body geometry and human pose estimation in video sequences, J. Vis. Commun. Image Represent. 82 (2022) 103407.

[27] M. Nasir, K. Muhammad, A. Ullah, J. Ahmad, S. Wook Baik, M. Sajjad, Enabling automation and edge intelligence over resource constraint IoT devices for smart home, Neurocomputing 491 (2022) 494–506.

[28] H. Zhang, W. Cui, T. Shi, Y. Tao, J. Zhang, ATMLP: Attention and time series MLP for fall detection, IAENG Int. J. Appl. Math. 53 (1) (2023) 1–8.

[29] Y.M. Galvão, J. Ferreira, V.A. Albuquerque, P. Barros, B.J. Fernandes, A multimodal approach using deep learning for fall detection, Expert Syst. Appl. 168 (2021) 114226.