A Lightweight Skeleton-Based Fall Detection Framework Using Multi-Dimensional Attention Mechanisms

Zongfei Zhang[1], Haoze Ni[2]
[1]Independent Researcher, Washington, United States
[2]College of Communication , Emerging Media Studies(EMS), Boston University, Boston, United States

*Corresponding author: zhangzongfei007@gmail.com

Abstract

Fall detection is a critical task in intelligent healthcare and smart monitoring systems, particularly for elderly care, where timely and reliable detection can significantly reduce the risk of severe injuries. In this paper, we propose a lightweight fall detection framework based on human skeleton sequences and multi-dimensional attention mechanisms. Instead of relying on raw RGB information, the proposed approach leverages skeleton-based representations to reduce sensitivity to background clutter, illumination variation, and appearance differences.

The framework consists of four main stages: skeleton keypoint extraction and preprocessing, spatial feature encoding using a Coordinate Attention enhanced Transformer, temporal motion modeling via a Temporal Attention mechanism, and final classification. The Coordinate Attention module enables direction-aware spatial modeling of human joints, while the Temporal Attention mechanism adaptively emphasizes critical motion phases associated with falls, such as sudden posture collapse and rapid descent.

Extensive experiments are conducted on two public benchmark datasets, namely the University of Rzeszow Fall Detection Dataset (URFD) and the Multiple Cameras Fall Dataset (MCFD). The experimental results demonstrate that the proposed method consistently outperforms several representative state-of-the-art fall detection approaches, achieving higher precision, recall, and F_1-score. In particular, the significant improvement in recall highlights the effectiveness of the proposed model in reducing missed fall detections, which is crucial for safety-critical applications.

Furthermore, the proposed framework is designed with computational efficiency in mind. By employing a compact Transformer architecture and lightweight attention modules, the system achieves real-time inference on CPU-based platforms, making it suitable for deployment in resource-constrained environments such as smart homes and elderly care facilities. Overall, this work demonstrates that combining skeleton-based representations with efficient spatial–temporal attention mechanisms provides a practical and reliable solution for real-world fall detection.

Index Terms— Fall detection, Skeleton-based representation, Human activity recognition, Pose estimation, Attention mechanisms, Transformer, Temporal attention, Lightweight model, Real-time monitoring.

1 Introduction

Falls are one of the leading causes of injury and hospitalization among the elderly population, posing a serious threat to personal safety and public healthcare systems[24]. With the rapid growth of aging societies worldwide, the demand for intelligent fall detection systems capable of continuous monitoring and timely response has increased significantly[7]. An effective fall detection system can provide early warnings and facilitate prompt assistance, thereby reducing the risk of severe injuries and mortality[22].

Traditional fall detection approaches often rely on wearable sensors such as accelerometers and gyroscopes[6]. Although these systems can achieve high detection accuracy, they require users to wear dedicated devices, which may cause discomfort and suffer from low compliance, especially among elderly individuals. Vision-based fall detection methods have therefore emerged as a promising alternative, as they enable non-contact monitoring and can be seamlessly integrated into existing surveillance infrastructures[10].

Early vision-based approaches primarily utilize raw RGB video data to extract appearance or motion features for fall detection[15]. However, these methods are highly sensitive to environmental factors such as illumination changes, background clutter, and camera viewpoints, and often incur high computational costs. To overcome these limitations, skeleton-based representations have gained increasing attention[11]. By modeling human motion using joint coordinates, skeleton-based methods effectively suppress irrelevant visual information and focus on posture dynamics and kinematic structures that are directly related to fall events.

Recent advances in pose estimation frameworks, such as MediaPipe Pose, have made reliable skeleton extraction from monocular RGB videos feasible in real time[17]. This has further promoted the adoption of skeleton-based fall detection models. Nevertheless, deep learning techniques, particularly attention-based models, have demonstrated strong capability

in modeling long-range dependencies and salient patterns in sequential data. Transformer architectures and attention mechanisms have been introduced to skeleton-based action recognition tasks, yielding promising results[20, 13]. However, vanilla Transformers often ignore directional spatial cues inherent in human motion and introduce substantial computational overhead, which restricts their applicability in real-time and resource-constrained scenarios.

To address these challenges, this paper proposes a lightweight fall detection framework based on human skeleton sequences and multi-dimensional attention mechanisms. The proposed approach integrates a Coordinate Attention enhanced Transformer to capture direction-aware spatial relationships among joints and a Temporal Attention mechanism to emphasize critical motion phases during falls. By jointly modeling spatial structure and temporal dynamics within a compact architecture, the proposed method achieves both high detection accuracy and real-time performance on CPU-based platforms.

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

- We propose a lightweight skeleton-based fall detection framework that effectively captures both spatial posture deformation and temporal motion dynamics using multi-dimensional attention mechanisms.
- A Coordinate Attention enhanced Transformer is introduced to improve spatial modeling of skeletal joints by incorporating direction-aware information.
- A Temporal Attention mechanism is employed to highlight key frames associated with abrupt posture transitions, significantly improving recall for fall events.
- Extensive experiments on public benchmark datasets demonstrate that the proposed method outperforms several state-of-the-art approaches while maintaining real-time inference capability.

The remainder of this paper is organized as follows. Section 2 reviews related work on fall detection and skeleton-based modeling. Section 3 describes the proposed methodology in detail. Section 4 presents experimental results and performance evaluations. Section 5 discusses the findings and limitations, and Section 6 concludes the paper.

2 Related Works

Fall detection has attracted extensive attention in computer vision and intelligent healthcare due to its critical role in elderly care and safety monitoring. This section reviews representative studies in vision-based methods, skeleton-based methods, and deep learning models with areas of attention mechanisms and highlights their limitations.

2.1 Vision-Based Fall Detection Methods

Early fall detection systems mainly relied on vision-based techniques using RGB video data. These methods typically extract appearance or motion features, such as optical flow, silhouettes, bounding box aspect ratios, and motion energy images, which are then classified using traditional machine learning models or convolutional neural networks (CNNs)[1, 4]. While such approaches have demonstrated promising performance in controlled environments, they are highly sensitive to illumination changes, background clutter, camera viewpoints, and occlusions. Moreover, processing raw video frames often incurs high computational cost, limiting real-time deployment on resource-constrained devices.

2.2 Skeleton-Based Fall Detection

To address the limitations of RGB-based approaches, skeleton-based fall detection methods have gained increasing popularity. By representing human motion using joint coordinates, skeleton-based models effectively suppress background noise and appearance variations, focusing instead on body posture and kinematic structure. With the advancement of pose estimation frameworks such as OpenPose and MediaPipe Pose, reliable skeleton extraction from monocular videos has become feasible[17, 3].

Several studies have utilized handcrafted skeletal features, including joint angles, body orientation, and center-of-mass trajectories, combined with classical classifiers for fall detection[5, 25, 9]. Although these methods are interpretable and computationally efficient, their performance is often limited by manual feature design and insufficient modeling of complex spatial dependencies among joints.

2.3 Deep Learning Models for Skeleton-Based Action Recognition

Inspired by progress in human action recognition, deep learning models have been introduced to automatically learn discriminative representations from skeleton sequences. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been employed to model temporal dynamics of joint movements. However, these models suffer from limitations such as gradient vanishing and limited parallelization capability[16, 12].

More recently, Transformer-based architectures have been explored for skeleton sequence modeling due to their ability to capture long-range dependencies through self-attention mechanisms[26, 21]. These models achieve strong performance in modeling global spatial and temporal relationships. Nevertheless, vanilla Transformers often overlook directional spatial cues inherent in human skeletal motion and introduce high computational overhead, which restricts their applicability in real-time fall detection systems.

2.4 Attention Mechanisms and Lightweight Designs

Attention mechanisms have been widely adopted to enhance the discriminative capability of fall detection models. Spatial attention focuses on anatomically important joints, while

temporal attention highlights critical motion phases associated with falls[8, 18]. Some works employ multi-level or hierarchical attention to improve robustness under complex motion patterns. Despite their effectiveness, many attention-based models rely on heavy network backbones or full self-attention, resulting in increased computational complexity.

In parallel, lightweight model design has emerged as an important research direction for practical fall detection[19, 23]. Techniques such as model pruning, reduced network depth, and efficient attention modules aim to balance detection accuracy with real-time performance. However, achieving high recall while maintaining low computational cost remains a challenging problem.

2.5 Positioning of This Work

Compared with existing methods, this work proposes a lightweight skeleton-based fall detection framework that jointly models spatial structure and temporal dynamics using multi-dimensional attention mechanisms. By integrating a Coordinate Attention enhanced Transformer for spatial encoding and a Temporal Attention module for motion emphasis, the proposed approach captures both anatomically meaningful joint dependencies and critical temporal transitions during falls. Furthermore, the compact architecture enables real-time inference on CPU-based platforms, making it suitable for practical deployment. This positions our method as an effective and deployable solution for real-world fall detection scenarios.

3 Methodology

In this study, we propose a lightweight fall detection model based on human skeleton sequences and multi-dimensional attention mechanisms. As illustrated in Fig. **??**, the framework consists of four key stages: (1) skeleton keypoint extraction, (2) spatial feature encoding via a Coordinate Attention (CA) enhanced Transformer, (3) Temporal Attention-based motion modeling, and (4) classification. The model focuses on capturing drastic motion changes during fall events and ensures real-time performance.

3.1 Skeleton Keypoint Extraction and Preprocessing

Accurate and robust human pose representation plays a crucial role in fall detection, as the dynamics of human posture during the falling process exhibit significant structural variations. In this study, the MediaPipe Pose framework is employed to extract 33 skeletal keypoints from each frame, covering the head, torso, and upper and lower limbs. For each keypoint i at time t, the model outputs a normalized spatial coordinate and a confidence score, represented as:

$$x_t^{(i)} = [u_t^{(i)}, v_t^{(i)}, c_t^{(i)}], \tag{1}$$

where $u_t^{(i)}$ and $v_t^{(i)}$ denote the pixel coordinates within the image plane after normalization, and $c_t^{(i)} \in [0, 1]$ indicates the reliability of the detection. With a sequence length of T, the input skeleton data can be formulated as:

$$X = \{x_t \mid t = 1, 2, \ldots, T\}, \quad x_t \in \mathbb{R}^{33\times 3}. \tag{2}$$

Although MediaPipe provides high-quality pose estimation, raw keypoints inevitably suffer from temporal jitter, occasional missing detections, and scale inconsistencies caused by the variability of human appearance and camera viewpoint. To ensure stability and spatial comparability across frames, a series of preprocessing operations are applied. First, a temporal median filter is performed on the coordinate trajectories to suppress random noise and reduce motion-induced fluctuations. When the confidence value of a keypoint is lower than a threshold, temporal interpolation is introduced to restore missing coordinates, effectively preventing invalid skeleton frames that may mislead subsequent learning.

Since the same action can be performed by subjects with different body sizes and positioned differently in the scene, the skeleton representation must be normalized to enhance generalization ability. To this end, the hip joint is designated as the reference origin for relocation while bone length normalization is applied to eliminate scale variations. The normalized form is expressed as:

$$x_t^{(i)} \leftarrow \frac{x_t^{(i)} - x_t^{(\text{hip})}}{\|x_t^{(\text{shoulder})} - x_t^{(\text{hip})}\|_2}, \tag{3}$$

where the denominator corresponds to the upper-body length, ensuring stable measurement even during movement. This transformation preserves the kinematic structure of the human body while enabling the model to focus on pose transitions rather than absolute positional displacement within the scene.

After preprocessing, the skeleton sequences are more consistent temporally and geometrically, providing reliable and noise-reduced motion cues for the subsequent spatial–temporal modeling. Such a refined representation greatly improves the robustness of fall detection under varying environments such as illumination variation, background clutter, and camera motion, making it highly suitable for real-world deployment scenarios.

3.2 Coordinate-Attention Enhanced Transformer

The spatial configuration and biomechanical constraints of the human skeleton play a decisive role in differentiating falls from other daily activities. During a fall, there is often a rapid collapse of torso posture accompanied by sudden positional changes in critical joints such as the head, hips, and knees. Therefore, effectively capturing the structural dependencies among skeleton joints is essential. In this work, we adopt a Transformer-based spatial encoding framework due to its global receptive capability and superior modeling of long-range interactions. For each frame, the preprocessed coordinates are concatenated and mapped into a latent embedding

vector through a learnable linear projection, followed by positional encoding to explicitly preserve temporal ordering:

$$p_t = W_{\text{emb}} \cdot \text{vec}(x_t) + b, \quad \tilde{p}_t = p_t + \text{PE}(t), \tag{4}$$

where $\text{PE}(t)$ is the standard sinusoidal positional encoder.

However, a vanilla Transformer only focuses on global token-to-token correlations and ignores directional spatial cues that are inherent in skeleton motion. To address this limitation, we embed a Coordinate Attention (CA) module [**?**] prior to Transformer encoding. Different from conventional channel attention, CA decomposes attention weights into horizontal and vertical dimensions, generating direction-aware representations that explicitly encode joint position distributions. This allows the network to emphasize anatomically significant regions and directional movement patterns, thus enhancing sensitivity to body collapse trajectories commonly observed in falls. The CA-enhanced feature representation is expressed as:

$$\hat{P} = \text{CA}(P), \tag{5}$$

where $P = [\tilde{p}_1, \ldots, \tilde{p}_T]$ denotes the original embedded sequence.

The enhanced skeleton features are then forwarded into a multi-layer Transformer encoder, where each layer consists of Multi-Head Self-Attention (MHSA) and a feed-forward subnetwork. The encoded spatio-structural representation can be formulated as:

$$H = \text{TransformerEncoder}(\hat{P}) \in \mathbb{R}^{T \times d}. \tag{6}$$

Through self-attention operations, joints that have large anatomical distances yet act synergistically during falls (e.g., head and hip positions) can establish direct attention connections, enabling the model to infer coordinated spatial collapse patterns. Benefiting from CA-enhanced position-aware interaction, the network achieves a more discriminative spatial understanding of pose deformation, paving the way for identifying key fall phases in later stages.

3.3 Temporal Attention Mechanism

A fall is not only characterized by distinct postures but also by the abrupt temporal transition from upright stability to rapid descent. Merely modeling spatial structures lacks the ability to emphasize these critical dynamic phases. To incorporate temporal discriminativeness, a Temporal Attention mechanism is introduced, which dynamically highlights frames that contribute more to fall identification.

Given the encoded sequence $H = \{h_1, h_2, \ldots, h_T\}$, temporal importance is learned through a trainable scoring function:

$$e_t = w^\top h_t, \tag{7}$$

and then normalized using a softmax operation to obtain temporal attention weights:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)}. \tag{8}$$

The final temporal-aware representation is obtained by a weighted summation over time:

$$z = \sum_{t=1}^{T} \alpha_t h_t. \tag{9}$$

This mechanism enables the network to automatically attend to crucial motion frames such as the initial loss of balance, sudden height descent, and post-fall ground contact posture. Frames representing steady walking or minor pose adjustments are suppressed, effectively reducing temporal redundancy. As falls often unfold in a short duration, the proposed attention model excels at capturing transient pose variability, leading to more accurate detection even in visually ambiguous scenarios. Moreover, the computational overhead introduced by this module is minimal, allowing the system to operate efficiently in real-time applications.

The combination of Transformer-based spatial encoding and temporal attention-driven dynamic emphasis establishes a comprehensive modeling framework that fully exploits the spatio-temporal properties of skeletal motion, thereby greatly improving the reliability of fall event recognition in diverse environments.

3.4 Classification and Training Objective

Following the spatio-temporal encoding process, the aggregated representation z is fed into a fully-connected classification layer to determine whether a fall has occurred. Since fall detection is formulated as a binary event recognition task comprising two classes — fall and non-fall — a Softmax activation is utilized to generate probabilistic outcomes:

$$\hat{y} = \text{Softmax}(W_c z + b_c), \tag{10}$$

where W_c and b_c denote learnable parameters. The output $\hat{y}$ represents the predicted likelihood for each class. To effectively optimize model parameters, a binary cross-entropy loss function is applied during training:

$$\mathcal{L} = -\sum_{c=1}^{2} y_c \log(\hat{y_c}), \tag{11}$$

where y_c is a one-hot encoded ground truth label. The objective of training is to minimize $\mathcal{L}$, enabling the model to differentiate abrupt body posture transitions associated with falls from other daily behaviors.

To further improve convergence stability and avoid overfitting, Adam optimizer is adopted with decoupled weight decay, accompanied by a cosine learning rate scheduling strategy. Early stopping is also employed based on validation performance to prevent unnecessary over-training. Through end-to-end optimization, the entire architecture learns a discriminative decision boundary that is highly responsive to fall-specific motion cues while maintaining robustness to inter-individual variations in movement patterns.

3.5 Lightweight Design for Real-Time Deployment

Considering practical deployment scenarios such as elderly care facilities and smart home monitoring systems, real-time performance and computational efficiency are critical. To ensure that the proposed method can operate smoothly on embedded and mobile platforms, several lightweight design strategies are incorporated into the architecture.

First, the Transformer encoder is designed in a compact configuration by reducing the number of attention heads and limiting the depth of stacked layers without sacrificing essential modeling capacity. This enables the network to maintain strong global spatial reasoning while keeping computational overhead low. Meanwhile, both the Coordinate Attention module and the Temporal Attention mechanism introduce minimal additional parameters, as they primarily rely on lightweight pooling and linear transformations. As a result, the model effectively improves spatial–temporal sensitivity without significantly increasing memory demand.

Moreover, redundant channels in the embedding dimension are pruned during training to reduce unnecessary computation in inference. Batch normalization fusion and matrix shape optimization further enhance speed and memory efficiency during deployment. These design considerations ensure that the model processes skeleton streams at frame-level granularity, enabling more than 30 frames per second on standard CPU-based edge devices.

In summary, the lightweight architecture not only preserves high recognition accuracy but also offers the practical advantage of real-time inference with low latency and low resource consumption. This makes the proposed system highly suitable for continuous monitoring applications, where timely and reliable detection of fall incidents plays an essential role in ensuring user safety.

4 Experiments

This section presents a comprehensive evaluation of the proposed fall detection method based on skeleton sequences and multi-dimensional attention mechanisms. We conduct experiments on benchmark datasets, compare our approach with several state-of-the-art models, perform ablation studies to validate the effectiveness of each module, and further analyze interpretability and real-time performance.

4.1 Datasets and Experimental Setup

To verify the generalization capability of the proposed method, we conduct experiments on two public benchmark fall detection datasets: University of Rzeszow Fall Detection Dataset (URFD) and Multiple Cameras Fall Dataset (MCFD). These datasets contain a wide variety of daily activities and fall scenarios recorded from different viewpoints, making them suitable for evaluating the discriminability of motion patterns.

For all sequences, we extract 33 human skeletal keypoints using MediaPipe Pose, forming fixed-length skeleton clips. Subjects are split into training and testing sets to ensure person-independent evaluation. The model is trained using the Adam optimizer with an initial learning rate of 10^{-4} for 100 epochs. Early stopping is applied based on validation loss to prevent overfitting. All experiments are conducted on a standard CPU environment (Intel i7), highlighting the lightweight and deployable nature of the proposed system.

Table 1: Performance evaluation results of each classifier on the test dataset

Paper	Characterization	Precision (%)	Recall (%)	$\mathbf{F_1}$ (%)
Asif et al. [2]	Segmentation and pose estimation	87.03	87.15	87.08
Yuan et al. [27]	Direction judgement and pose estimation	88.13	86.66	87.38
Feng et al. [8]	Person detection, tracking and CNN	**89.80**	83.50	86.50
PIFR [14]	Pose angle and pose estimation	88.80	**94.10**	**91.40**
Proposed Model	Pose estimation	94.72	**98.00**	**96.33**

4.2 Performance Comparison with Existing Methods

Table 1 presents the comparative performance results of the proposed model against several state-of-the-art methods, including those of Asif et al. [2], Yuan et al. [27], Feng et al. [8], and Kong et al. (PIFR) [14]. The proposed model achieved the highest $\mathbf{F_1}$-score (96.33%), outperforming all baselines.

In contrast, PIFR [14] improved recall to 94.10% through pose-angle feature enhancement but exhibited slightly lower precision (88.80%).

The proposed model maintains a balanced performance, achieving 94.72% precision and 98.00% recall, demonstrating superior fall recognition ability with minimal false detections. This improvement stems from the model's *spatio-temporal attention mechanism*, which effectively captures both directional joint dependencies and abrupt motion transitions, thereby enabling a more discriminative representation of fall patterns.

The substantial improvement in recall indicates that the proposed approach is particularly effective at reducing missed detections, which is critical for safety-sensitive applications such as fall monitoring. Meanwhile, the consistently high precision demonstrates that the model maintains strong robustness against false alarms. Overall, these results confirm that the proposed model outperforms existing methods in terms of balanced detection accuracy and reliability, making it well suited for real-world fall detection scenarios.

5 Discussion

This study presents a lightweight fall detection framework that leverages skeleton-based representations and multi-dimensional attention mechanisms to achieve accurate and real-time performance. The experimental results demonstrate that the proposed model consistently outperforms several representative vision-based fall detection methods in terms of precision, recall, and F_1-score. In this section, we further discuss the key factors contributing to the observed performance gains, the advantages of the proposed design, and its limitations.

5.1 Effectiveness of Skeleton-Based Representation

One of the main advantages of the proposed approach lies in the use of human skeleton sequences rather than raw RGB data. Skeleton-based representations inherently suppress background clutter, illumination variations, and appearance differences among subjects, allowing the model to focus on motion dynamics and body posture changes that are directly related to fall events. This property is particularly beneficial in real-world environments where visual conditions are often uncontrollable. The experimental results indicate that such representations provide a more robust basis for fall detection compared to methods that rely heavily on pixel-level information or handcrafted visual features.

5.2 Impact of Spatial and Temporal Attention Mechanisms

The integration of a Coordinate Attention (CA) enhanced Transformer and a Temporal Attention mechanism plays a critical role in improving detection performance. The CA module enables the model to capture direction-aware spatial dependencies among key joints, which is essential for recognizing characteristic body collapse patterns during falls. Meanwhile, the Temporal Attention mechanism adaptively emphasizes frames corresponding to abrupt posture transitions, such as loss of balance and ground contact, while suppressing redundant or non-informative frames.

The superior recall achieved by the proposed model suggests that the attention-driven design is particularly effective in reducing missed detections. This is a crucial property for safety-critical applications, where failing to detect a fall may lead to severe consequences. At the same time, the high precision indicates that the model maintains strong discrimination capability against non-fall activities, avoiding excessive false alarms.

5.3 Lightweight Design and Practical Deployment

Another important aspect of this work is its emphasis on computational efficiency. By adopting a compact Transformer configuration and lightweight attention modules, the proposed system achieves real-time inference on standard CPU-based platforms without sacrificing detection accuracy. This makes the model suitable for deployment in resource-constrained environments such as smart homes, elderly care facilities, and edge devices, where low latency and energy efficiency are essential.

Compared with deep learning models that require heavy convolutional backbones or GPU acceleration, the proposed approach strikes a favorable balance between performance and computational cost. This balance is particularly important for continuous monitoring scenarios, where long-term operation and scalability are key considerations.

5.4 Limitations and Future Work

Despite its promising performance, the proposed method still has several limitations. First, the reliance on accurate skeleton extraction means that severe occlusions or extreme camera viewpoints may affect detection reliability. Second, although experiments are conducted on two public benchmark datasets, real-world environments may exhibit more complex interactions and variations that are not fully represented in the datasets.

Future work will focus on enhancing robustness under occlusion conditions, incorporating multi-view or multi-modal information, and extending the framework to detect other abnormal behaviors beyond falls. In addition, further evaluation on larger-scale real-world datasets will be conducted to validate long-term stability and generalization performance.

Overall, the results and analysis demonstrate that the proposed lightweight, attention-driven skeleton-based framework offers a reliable and practical solution for fall detection, bridging the gap between high detection accuracy and real-time deployability.

6 Conclusion

In this paper, we presented a lightweight and effective fall detection framework based on human skeleton sequences and multi-dimensional attention mechanisms. By integrating a Coordinate Attention enhanced Transformer for spatial modeling and a Temporal Attention mechanism for dynamic motion emphasis, the proposed method is able to capture both structural posture deformation and abrupt temporal transitions that are characteristic of fall events.

Extensive experiments conducted on two public benchmark datasets, namely the University of Rzeszow Fall Detection Dataset (URFD) and the Multiple Cameras Fall Dataset (MCFD), demonstrate that the proposed model outperforms several representative state-of-the-art methods in terms of precision, recall, and F_1-score. In particular, the significant improvement in recall indicates the model's strong capability in reducing missed fall detections, which is crucial for safety-critical monitoring applications. At the same time, the high precision confirms its robustness against false alarms caused by daily activities.

In addition to detection accuracy, the proposed framework emphasizes computational efficiency and real-time performance. Through a compact Transformer design and lightweight attention modules, the model achieves real-time inference on CPU-based platforms, making it suitable for deployment in resource-constrained environments such as smart homes and elderly care facilities.

Overall, this work demonstrates that combining skeleton-based representations with carefully designed spatial–temporal attention mechanisms offers a promising solution for practical fall detection systems. Future work will focus on improving robustness under occlusion and complex environmental conditions, as well as extending the framework to broader abnormal behavior recognition tasks.

References

[1] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif. Activity recognition for indoor fall detection using convolutional neural network. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 81–84. IEEE, 2017.

[2] Umar Asif, Benjamin Mashford, Stefan Von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, and Stefan Harrer. Privacy preserving human fall detection using video data. In *Machine Learning for Health Workshop*, pages 39–51. PMLR, 2020.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[4] Sarah Almeida Carneiro, Gabriel Pellegrino da Silva, Guilherme Vieira Leite, Ricardo Moreno, Silvio Jamil F Guimaraes, and Helio Pedrini. Multi-stream deep convolutional network using high-level features applied to fall detection in video sequences. In *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 293–298. IEEE, 2019.

[5] Weiming Chen, Zijie Jiang, Hailin Guo, and Xiaoyang Ni. Fall detection based on key points of human-skeleton using openpose. *Symmetry*, 12(5):744, 2020.

[6] Yueng Santiago Delahoz and Miguel Angel Labrador. Survey on fall detection and fall prevention using wearable and external sensors. *Sensors*, 14(10):19806–19842, 2014.

[7] Nashwa El-Bendary, Qing Tan, Frédérique C Pivot, and Anthony Lam. Fall detection and prevention for the elderly: A review of trends and challenges. *International Journal on Smart Sensing and Intelligent Systems*, 6(3):1230, 2013.

[8] Qi Feng, Chenqiang Gao, Lan Wang, Yue Zhao, Tiecheng Song, and Qiang Li. Spatio-temporal fall event detection in complex scenes using attention guided lstm. *Pattern Recognition Letters*, 130:242–249, 2020.

[9] Martha Magali Flores-Barranco, Mario-Alberto Ibarra-Mazano, and Irene Cheng. Accidental fall detection based on skeleton joint correlation and activity boundary. In *International Symposium on Visual Computing*, pages 489–498. Springer, 2015.

[10] Jesús Gutiérrez, Víctor Rodríguez, and Sergio Martin. Comprehensive review of vision-based fall detection systems. *Sensors*, 21(3):947, 2021.

[11] Van-Ha Hoang, Jong Weon Lee, Md Jalil Piran, and Chun-Su Park. Advances in skeleton-based fall detection in rgb videos: From handcrafted to deep learning approaches. *IEEe Access*, 11:92322–92352, 2023.

[12] Anitha Rani Inturi, VM Manikandan, and Vignesh Garrapally. A novel vision-based fall detection scheme using keypoints of human skeleton with long short-term memory network. *Arabian Journal for Science and Engineering*, 48(2):1143–1155, 2023.

[13] Duncan Kibet, Min Seop So, Hahyeon Kang, Yongsu Han, and Jong-Ho Shin. Sudden fall detection of human body using transformer model. *Sensors*, 24(24):8051, 2024.

[14] Vungsovanreach Kong, Saravit Soeng, Munirot Thon, Wan-Sup Cho, Anand Nayyar, and Tae-Kyung Kim. Pifr: A novel approach for analyzing pose angle-based human activity to automate fall detection in videos. *PLoS One*, 20(6):e0325253, 2025.

[15] Sergio Lafuente-Arroyo, Pilar Martín-Martín, Cristian Iglesias-Iglesias, Saturnino Maldonado-Bascón, and Francisco Javier Acevedo-Rodríguez. Rgb camera-based fallen person detection system embedded on a mobile platform. *Expert Systems with Applications*, 197:116715, 2022.

[16] Chuan-Bi Lin, Ziqian Dong, Wei-Kai Kuan, and Yung-Fa Huang. A framework for fall detection based on openpose skeleton and lstm/gru models. *Applied Sciences*, 11(1):329, 2020.

[17] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[18] Haoze Ni, Xinyue Huang, and Wuyang Zhang. From detection to prediction: A multimodal deep learning framework for proactive fall risk monitoring in smart aging. *INNO-PRESS: Journal of Emerging Applied AI*, 1(6), 2025.

[19] Nadhira Noor and In Kyu Park. A lightweight skeleton-based 3d-cnn for real-time fall detection and action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2179–2188, 2023.

[20] Adrián Núñez-Marcos and Ignacio Arganda-Carreras. Transformer-based fall detection in videos. *Engineering Applications of Artificial Intelligence*, 132:107937, 2024.

[21] Ali Raza, Muhammad Haroon Yousaf, Sergio A Velastin, and Serestina Viriri. Human fall detection from sequences of skeleton features using vision transformer. In *VISIGRAPP (5: VISAPP)*, pages 591–598, 2023.

[22] Lingmei Ren and Yanjun Peng. Research of fall detection and fall prevention technologies: A systematic review. *IEEE access*, 7:77702–77722, 2019.

[23] Yuyang Sha, Xiaobing Zhai, Junrong Li, Weiyu Meng, Henry HY Tong, and Kefeng Li. A novel lightweight deep learning fall detection system based on global-local attention and channel feature augmentation. *Interdisciplinary Nursing Research*, 2(2):68–75, 2023.

[24] Raju Vaishya and Abhishek Vaish. Falls in older adults are serious. *Indian journal of orthopaedics*, 54(1):69–74, 2020.

[25] Leiyue Yao, Weidong Min, and Keqiang Lu. A new approach to fall detection based on the human torso motion model. *Applied Sciences*, 7(10):993, 2017.

[26] Xiaoqun Yu, Chenfeng Wang, Wenyu Wu, and Shuping Xiong. A real-time skeleton-based fall detection algorithm based on temporal convolutional networks and transformer encoder. *Pervasive and Mobile Computing*, 107:102016, 2025.

[27] Chunmiao Yuan, Pengju Zhang, Qingyong Yang, and Jianming Wang. Fall detection and direction judgment based on posture estimation. *Discrete dynamics in nature and society*, 2022(1):8372291, 2022.