RESEARCH ARTICLE

Lightweight Cascaded Feature Reweighting for Fall Detection through Context-Aware YOLOv8 Architecture

Farhan Ali[1], **Alexandros Gazis**[2,3] and **Faryal Zahoor**[4,*]

[1] Faculty of Computer Science and Biomedical Engineering, Graz University of Technology, 8010 Graz, Austria
[2] Department of Computer Science, School of Sciences, Democritus University of Thrace, 65404 Kavala, Greece
[3] Heriot-Watt University, Edinburgh Campus, Edinburgh EH14 4AS, United Kingdom
[4] BRAINS Institute Peshawar, Peshawar 25000, Pakistan

Abstract

Falls represent a significant global health concern, particularly among older adults, with delayed detection often leading to severe medical complications. Although computer vision-based fall detection systems offer promising solutions, they usually struggle with diverse real-world scenarios and computational efficiency. This paper introduces a novel lightweight cascaded feature reweighting approach that enhances YOLOv8 for reliable fall detection through a context-aware architecture. We strategically integrate three complementary attention mechanisms: Squeeze-and-Excitation blocks in the early stages, Spatial Attention modules in the later stages, and Efficient Channel Attention in the neck section, creating a progressive feature refinement pipeline that leverages the bilateral symmetry properties of human posture. Our approach achieves significant performance improvements, with an mAP of 0.878, an increase of 1.5% over the baseline YOLOv8, while maintaining minimal computational overhead. This makes it well-suited for real-world deployment in resource-constrained environments. Comprehensive evaluations across two diverse datasets, including DiverseFall and CAUCAFall, demonstrate that our model outperforms state-of-the-art (SOTA) detectors, including Faster R-CNN and earlier YOLO variants. Our approach shows particularly pronounced advantages under challenging and varied environmental conditions. Ablation studies confirm the effectiveness of our architectural design choices, demonstrating that each attention mechanism makes a unique contribution to overall performance improvement. The proposed lightweight architecture represents a significant advancement in vision-based fall detection, striking a balance between high accuracy and computational efficiency while maintaining robust performance in diverse real-world scenarios.

Keywords: fall detection, lightweight architecture, feature reweighting, healthcare monitoring, elderly care, elderly fall detection, computer vision in healthcare.

Academic Editor:
Xue-Bo Jin

Submitted: 22 May 2025
Accepted: 03 July 2025
Published: 06 November 2025

Vol. 2, **No.** 4, 2025.
10.62762/TIS.2025.196437

***Corresponding author:**
✉ Faryal Zahoor
faryal.zahoorjan@gmail.com

Citation
Ali, F., Gazis, A., & Zahoor, F. (2025). Lightweight Cascaded Feature Reweighting for Fall Detection through Context-Aware YOLOv8 Architecture. *ICCK Transactions on Intelligent Systematics*, 2(4), 224–237.

1 Introduction

Falls represent a significant public health concern throughout the world, with potentially devastating consequences ranging from minor injuries to fatalities. According to the World Health Organization (WHO), falls result in approximately 684,000 deaths annually, with adults aged 60 years and older experiencing the highest mortality rates [1]. The impact of falls extends beyond physical harm, affecting psychological well-being and quality of life, particularly among older adults. Undetected falls are especially dangerous for elderly people and those with limited mobility, where delayed assistance can transform a manageable incident into a complex medical emergency, increasing the burden of healthcare and associated costs. Human posture typically exhibits bilateral symmetry, a property that can be leveraged to enhance fall detection accuracy in complex environments and with occlusions. However, traditional detection methods often struggle to capitalize on these inherent characteristics, highlighting the need for more sophisticated approaches. Given these challenges, there is an urgent requirement for reliable, intelligent fall detection (FD) systems that can improve individual safety, particularly in healthcare and assisted living settings. Existing FD systems can be categorized into three primary groups [2]: ambient sensor-based approaches, wearable sensor-based approaches, and computer vision (CV)-based approaches. Ambient sensor systems employ environmental monitoring devices to collect data on parameters such as pressure, vibration, and sound. However, these systems offer limited spatial coverage due to their fixed installation locations and often lack contextual awareness, compromising detection accuracy and reliability [3]. Wearable sensor solutions require users to wear devices equipped with accelerometers, gyroscopes, and magnetometers in various body locations (e.g., chest, back, or waist) to monitor motion patterns for fall detection [4]. Despite their effectiveness, these systems present significant usability challenges. Elderly individuals may find them uncomfortable or forget to wear them consistently, while battery limitations —though necessary— require regular charging or replacement, often causing additional inconvenience [5].

Vision-based FD systems using fixed cameras have gained prominence due to their passive monitoring capabilities, continuous power supply, and the elimination of wearable equipment requirements [6]. Deep learning (DL)-based vision systems, particularly those utilizing RGB or depth cameras (e.g., CCTV systems), have emerged as powerful tools for fall detection [7–11]. Multimodal approaches that combine camera and sensor data [12–15] offer enhanced performance but introduce complexity in data integration and interpretation, thereby complicating practical deployment [12]. Among deep learning (DL)-based methods, the You Only Look Once (YOLO) family of algorithms has demonstrated superior performance in object detection across various domains, thanks to its exceptional speed and accuracy [16, 17]. Previous research has explored various YOLO implementations for fall detection: YOLOv2 with pre-trained CNNs on the MS-COCO dataset [18], YOLOv3 combined with Support Vector Machines (SVMs) for posture recognition, a fall management system using monocular cameras and humanoid robots [19], and a YOLOv4-based approach utilizing the UR Fall Detection dataset [11]. Despite these advancements, many existing FD models remain limited by the insufficient diversity of training data, which constrains their generalizability and scalability. For instance, while [20] introduced a YOLOv7-fall model for rapid detection, their dataset lacked diversity, reducing the model's effectiveness across varied scenarios. To address these limitations, our research focuses on developing a vision-based state-of-the-art (SOTA) fall detection approach. The technical merit of this article is demonstrated through the following key contributions:

- A robust and scalable fall detection framework that integrates strategically placed attention mechanisms to progressively enhance feature learning. Our design ensures improved generalizability across diverse real-world scenarios, addressing a key limitation of many existing fall detection systems.

- Improved detection accuracy via targeted YOLO architecture modifications, achieved through the integration of three complementary attention modules: Squeeze-and-Excitation (SE) blocks for early-stage channel recalibration, Spatial Attention (SA) blocks for later-stage regional focus, and Efficient Channel Attention (ECA) blocks in the neck for refined feature fusion. This cascaded feature reweighting strategy substantially strengthens the model's ability to capture fall-relevant features.

- High generalizability with parameter efficiency, delivering substantial performance gains

while adding only 0.2M parameters to the baseline YOLOv8n model. The resulting lightweight design enables real-time inference on resource-constrained devices without compromising accuracy in diverse environmental conditions.

- A comprehensive evaluation framework that rigorously assesses fall detection performance through detailed ablation studies and comparative analyses against state-of-the-art detectors. These evaluations demonstrate the model's superior accuracy and robustness across multiple metrics, datasets, and challenging real-world conditions.

The rest of this paper is organized as follows: Section 2 reviews related work on fall detection systems, Section 3 describes our proposed methodology, Section 4 presents experimental results and analysis, and Section 5 concludes with a discussion of limitations and future research directions.

2 Related Work

Fall detection (FD) research has evolved primarily along two trajectories: traditional sensor-based approaches and more recent vision-based methods. Table 1 summarizes key methodologies from various fall detection studies. This section examines these approaches, highlighting their respective strengths and limitations.

2.1 Sensor-Based Fall Detection

Traditional sensor-based fall detection systems employ machine learning (ML) algorithms to analyze data from wearable devices or ambient sensors. Kwolek et al. [27] combined visual frame data with Support Vector Machine (SVM) classifiers for effective fall detection. In a different approach, Yacchirema et al. [28] integrated accelerometer technology with wearable devices using decision tree algorithms to automatically alert caregivers upon fall detection. Several researchers have explored alternative sensing modalities to improve detection accuracy. For instance, Seredin et al. [29] developed a privacy-preserving approach utilizing skeletal feature encoding with SVM classification. Meanwhile, Chen et al. [30] analyzed accelerometer data from wristwatches, though hand movement interference remains a significant challenge with this approach. Despite their practical applications, sensor-based approaches suffer from notable limitations, including poor user compliance with wearable devices, restricted monitoring range with ambient sensors, and difficulty capturing the contextual information necessary for accurate fall detection in real-world scenarios.

2.2 Vision-Based Fall Detection

Recent advances in deep learning have shifted the focus toward vision-based fall detection systems, with YOLO (You Only Look Once) architectures gaining prominence due to their real-time performance capabilities. The progressive evolution of YOLO variants has yielded increasingly sophisticated fall detection systems tailored to this specific use case. Early implementations utilized YOLOv2 [18] for human detection with pre-trained weights, further fine-tuned on manually annotated fall images. Building on this foundation, Raza et al. [11] developed a YOLOv4-based network trained on a dataset containing approximately 1,691 fall and 1,731 normal samples, demonstrating the ability to recognize falls using standard visual sensors without requiring additional environmental sensors. More recent research has focused on architectural enhancements to the YOLO framework specifically for fall detection scenarios. Chen et al. [31] modified YOLOv5s by replacing conventional convolutions with asymmetric convolution blocks and incorporating spatial attention mechanisms to improve feature extraction capabilities. Meanwhile, Zhao et al. [20] introduced YOLOv7-fall, claiming improved feature extraction with fewer model parameters, although the training dataset was relatively limited in size and diversity.

Despite recent advancements, vision-based approaches still face persistent challenges. Many current implementations lack architectural optimizations tailored to the specific demands of fall detection. Furthermore, most studies rely on relatively homogeneous datasets, which restrict model generalization to real-world environments characterized by diverse lighting conditions, camera angles, and occlusions. These limitations underscore the need for architectural innovations specifically designed for fall detection, as well as evaluation on more diverse datasets to enhance real-world applicability. Our work addresses these gaps by introducing targeted modifications to the YOLO architecture and employing a more comprehensive and diverse training dataset.

Table 1. Comparative analysis of fall detection techniques and datasets from recent literature.

Year	Technique	Dataset	Sensors
2019 [5]	Body posture angle, SVM	Real-time data	MPU6500 sensor
2020 [34]	CNN and SVM	FPDS, SCDS	RGB camera
2021 [15]	Multimodal CNN	UR Fall, UP-Fall	RGB images, accelerometers
2022 [31]	Modified YOLOv5s	URFD dataset	Microsoft Kinect cameras
2023 [32]	YOLOv5x, YOLOv5s	CAUCA Fall	Webcam, IoT devices
2024 [20]	YOLOv7-fall, YOLOv7-tiny	Multi-camera FD, UR FD	RGB cameras
2024 [33]	YOLOv8	DiverseFall	RGB cameras

3 Methodology

This section outlines our proposed fall detection approach, which enhances the YOLOv8 architecture by strategically incorporating attention mechanisms. We first provide an overview of the base YOLOv8 architecture, followed by detailed descriptions of the three attention mechanisms employed: Squeeze-and-Excitation (SE) blocks, Spatial Attention (SA) blocks, and Efficient Channel Attention (ECA) blocks. The complete architectural overview is illustrated in Figure 1.

3.1 YOLOv8 Architecture

Numerous YOLO-based and attention models have been explored to tackle various detection challenges, demonstrating the growth and flexibility [21, 22]. YOLOv8 marks a major step forward in real-time object detection frameworks, building on the success of earlier versions through architectural improvements that boost both accuracy and efficiency. The network features a three-part design consisting of a backbone, a neck, and a detection head. The backbone, based on CSPDarknet, extracts features from input images by using a series of convolutional layers arranged into stages that grow more semantically rich while reducing spatial resolution. The stem layer begins with a Focus module that efficiently consolidates information from a 2×2 spatial neighborhood into the channel dimension, reducing computation while retaining information. After the stem layer, the backbone has four main stages, each containing Cross Stage Partial Fusion (C2F) blocks, followed by convolutional layers. The C2F blocks use an efficient design that processes feature maps in parallel paths, improving information flow and gradient propagation. These stages progressively downsample the spatial dimensions while increasing the feature channels, generating multi-scale feature maps (P2 through P5) with varying resolutions ($80 \times 80 \times 256$ to $20 \times 20 \times 1024$).

The neck component functions as a feature fusion mechanism, aggregating outputs from various backbone stages to enhance both semantic richness and spatial detail. It leverages a Feature Pyramid Network (FPN) structure, combined with Path Aggregation Network (PAN) principles, to facilitate bidirectional information flow across different feature scales. This design incorporates upsampling operations to merge high-level semantic features with low-level spatial features, thereby improving object detection performance across a wide range of scales. The detection head receives these fused features and performs the final prediction tasks. It is composed of three parallel branches, each dedicated to detecting objects at a specific scale. Each branch outputs bounding box coordinates, objectness scores, and class probabilities through a sequence of convolutional operations. In our enhanced architecture, we strategically integrate three complementary attention mechanisms at distinct locations within the network to further improve fall detection accuracy. These modifications are described in detail in the following subsections.

3.2 Squeeze-and-Excitation (SE) Block

The Squeeze-and-Excitation (SE) block, integrated into Stages 1 and 2 of our architecture, aims to enhance feature representation by explicitly modeling interdependencies between channels. This lightweight attention mechanism adaptively recalibrates channel-wise feature responses, allowing the network to emphasize informative features while suppressing less useful ones. The SE block operates through two key operations: squeeze and excitation. The squeeze operation generates a channel descriptor by aggregating feature maps across their spatial dimensions using global average pooling. This descriptor captures the global distribution of channel-wise feature responses, producing a vector with size C, where C is the number of channels.

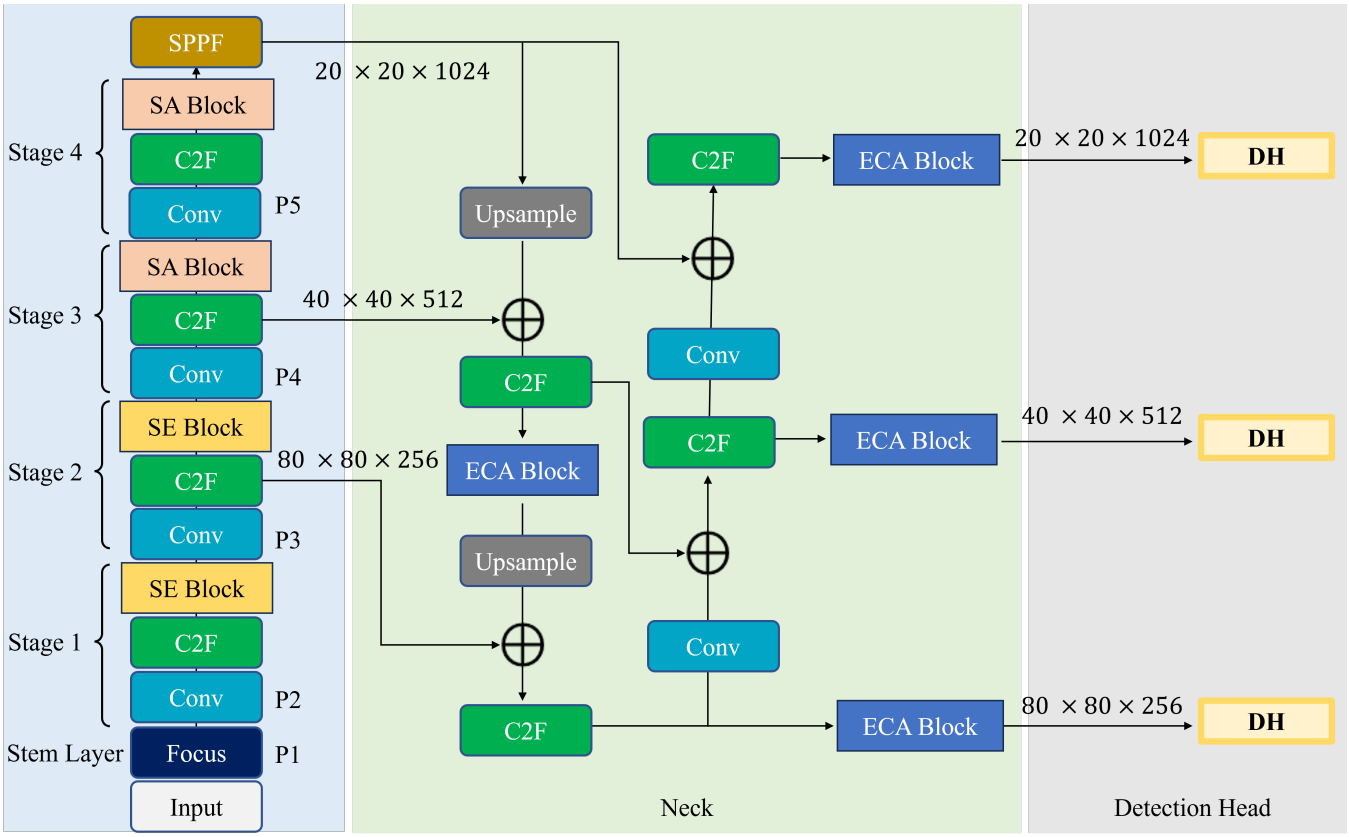

Figure 1. Visual overview of the entire model architecture. From left to right: (1) Backbone feature extraction (2) Neck (3) Detection Heads.

Formally, this is defined as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{1}$$

The excitation operation follows, employing a simple gating mechanism with two fully connected layers and a ReLU activation in between, followed by a sigmoid activation. This creates a set of channel-specific weights that are used to scale the original feature maps:

$$s = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where δ refers to the ReLU function, σ is the sigmoid activation, and $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are parameters of the fully connected layers with a reduction ratio r (typically set to 16). The final output of the SE block is obtained by rescaling the input feature map with the activated channel weights:

$$\tilde{x}_c = s_c \cdot x_c \tag{3}$$

In our fall detection architecture, SE blocks are specifically placed after the C2F modules in Stages 1 and 2. This strategic placement enables the network to focus on relevant channel information early in the feature extraction process, enhancing the low-level features that are crucial for distinguishing subtle posture changes and movement patterns characteristic of falls. The computational overhead of SE blocks is minimal, making them ideal for the early stages of the network, where maintaining efficiency is critical.

3.3 Spatial Attention Block

Attention mechanisms have been widely explored in computer vision to enhance feature representation and localization capabilities [23–25]. While channel attention mechanisms, such as SE blocks, focus on "what" is meaningful in the feature maps, Spatial Attention (SA) blocks emphasize "where" to focus, making them particularly valuable for fall detection, where spatial configurations are crucial. We integrate SA blocks into Stages 3 and 4 of our architecture to enhance the network's ability to focus on spatially relevant regions during fall events. The SA block generates a spatial attention map that highlights important areas of the feature maps. The process begins by aggregating channel information through two parallel operations: average pooling and max pooling along the channel axis. This generates two

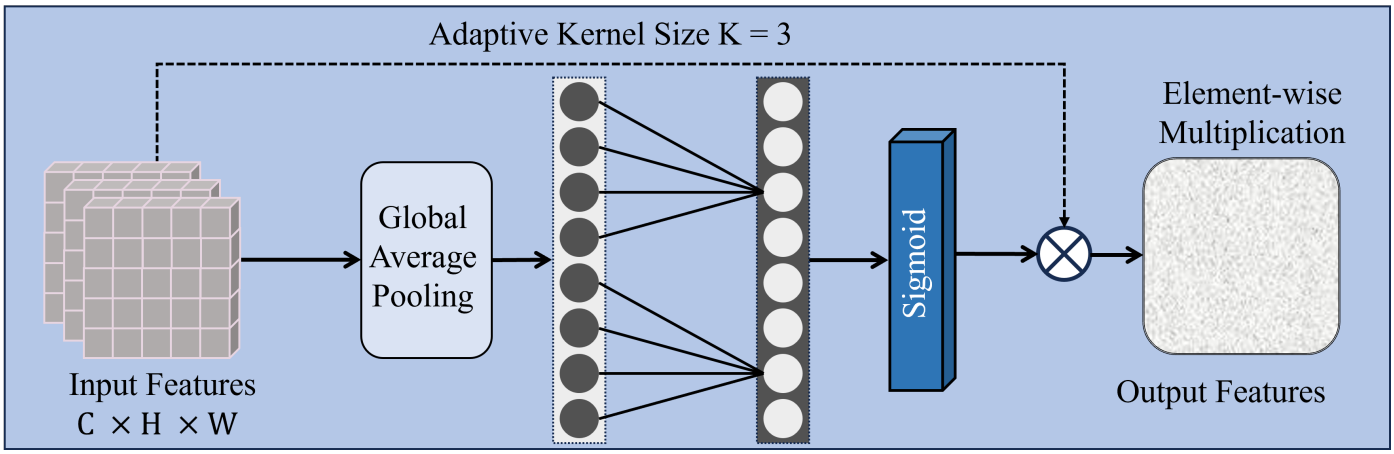

Figure 2. Visual overview of the Efficient Channel Attention (ECA) module.

2D spatial feature descriptors that capture different aspects of the spatial information:

$$F^s_{avg} = \frac{1}{C}\sum_{c=1}^{C} x_c \tag{4}$$

$$F^s_{max} = \max_{c=1}^{C} x_c \tag{5}$$

These pooled features are concatenated along the channel dimension and processed by a standard convolutional layer with a kernel size of 7×7 to generate the spatial attention map:

$$M_s = \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}])) \tag{6}$$

where σ denotes the sigmoid function, $f^{7\times7}$ represents a convolution operation with a 7×7 kernel, and $[F^s_{avg}; F^s_{max}]$ denotes channel-wise concatenation. The final output is obtained by performing element-wise multiplication between the spatial attention map and the input feature map:

$$\tilde{x} = M_s \otimes x \tag{7}$$

The SA blocks are placed after the C2F modules in Stages 3 and 4 of our architecture. This placement is intentional, as the higher-level feature maps in these stages contain more semantic information about object shapes and positions. By focusing attention on spatially relevant regions in these later stages, the network can better identify the distinctive spatial patterns associated with fall incidents, such as sudden changes in body orientation or position. This is particularly important for distinguishing falls from similar activities such as sitting or lying down, where the spatial configuration of the human body is the key differentiating factor.

3.4 Efficient Channel Attention (ECA) Block

The Efficient Channel Attention (ECA) block represents an improvement over traditional channel attention mechanisms by avoiding dimension reduction operations that can lead to information loss [26]. In our fall detection architecture, ECA blocks (as shown in Figure 2) are integrated throughout the neck section and at a key position in Stage 2, serving as efficient feature enhancers that preserve detailed information crucial for accurate fall detection. Unlike the SE block that employs fully connected layers for channel interaction, the ECA block captures local cross-channel interactions through a one-dimensional convolution operation. The process begins with global average pooling to generate channel descriptors:

$$y_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \tag{8}$$

Instead of using dimension-reducing fully connected layers, ECA applies a 1D convolution with a kernel size of k directly on these channel descriptors:

$$w_c = \sigma(\text{Conv1D}_k(y_c)) \tag{9}$$

where Conv1D_k represents a one-dimensional convolution with kernel size k. This operation allows each channel to interact with its k neighboring channels, capturing local cross-channel interactions without dimension reduction. The final output is obtained by rescaling the original feature maps with the attention weights:

$$\tilde{x}_c = w_c \cdot x_c \tag{10}$$

The kernel size k in ECA is adaptively determined based on the channel dimension to maintain a proper

coverage of channel interactions:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}} \tag{11}$$

where γ and b are hyperparameters, and $|\cdot|_{\text{odd}}$ ensures the result is an odd integer. In our architecture, ECA blocks are strategically placed at multiple locations: after the C2F blocks in the neck section and at a key position in Stage 2. This placement enables the ECA blocks to enhance feature representations at critical fusion points, where maintaining detailed information is crucial. Unlike traditional channel attention mechanisms, ECA preserves the original feature dimensions while adding minimal computational overhead, making it ideal for the neck section where feature fusion occurs. The preservation of detailed information is particularly valuable for fall detection, where fine-grained motion patterns and posture changes must be accurately captured for reliable detection.

3.5 Integration Strategy

Our proposed architecture strategically combines these three attention mechanisms to develop a progressive attention pattern throughout the network. This combination follows a carefully planned approach to maximize the strengths of each attention mechanism.

1. **Early Stages (1–2):** SE blocks are deployed to recalibrate channel information with minimal computational overhead, emphasizing important feature channels in the early feature extraction process.
2. **Later Stages (3–4):** SA blocks are implemented to focus on spatially relevant regions, enhancing the network's ability to identify the spatial patterns characteristic of falls.
3. **Neck and Fusion Points:** ECA blocks are placed at feature fusion locations to maintain and enhance detailed information without dimension reduction, ensuring that critical features for fall detection are preserved during the fusion process.

This progressive attention pattern creates a complementary system where channel attention dominates initially, spatial attention takes over in later stages, and efficient feature enhancement happens at fusion points. The result is a comprehensive attention mechanism that tackles the unique challenges of fall detection: detecting subtle posture changes, recognizing characteristic movement patterns, and preserving detailed feature information for accurate classification. Compared to using a single attention mechanism throughout the network, this strategic integration allows the architecture to leverage the strengths of each mechanism while reducing their limitations. The overall computational overhead remains low, ensuring that the improved architecture maintains the real-time performance needed for practical fall detection applications.

4 Results and Discussion

This section presents a comprehensive analysis of the proposed fall detection model, emphasizing its performance, efficiency, and the contribution of each attention mechanism. We evaluate the model using both quantitative metrics and qualitative insights to validate its effectiveness.

4.1 Experimental Setup

The proposed model was implemented in PyTorch and trained using a workstation equipped with an NVIDIA RTX 4090 GPU (24GB VRAM) and an Intel Core i9-10900X CPU. The following hyperparameters were used:

- **Optimizer:** Stochastic Gradient Descent (SGD) with weight decay of 5e-4.
- **Learning Rate:** Initialized at 0.001 with cosine annealing schedule.
- **Batch Size:** 16
- **Training Epochs:** 150.

4.2 Datasets and Preprocessing

Our model was trained and evaluated on two datasets: DiverseFALL10500 and CAUCAFall (Figure 3). For the DiverseFALL10500 dataset, we performed a stratified split with 70% for training, 20% for validation, and 10% for testing. This ensured class balance across all sets. Similar preprocessing steps and input resolutions were applied to the CAUCAFall datasets to ensure fair comparisons.

4.3 Evaluation Metrics

To evaluate the model's performance, we employed a comprehensive set of metrics, as detailed below.

4.3.1 Classification Metrics

We used Precision, Recall, and F1-score, defined as:

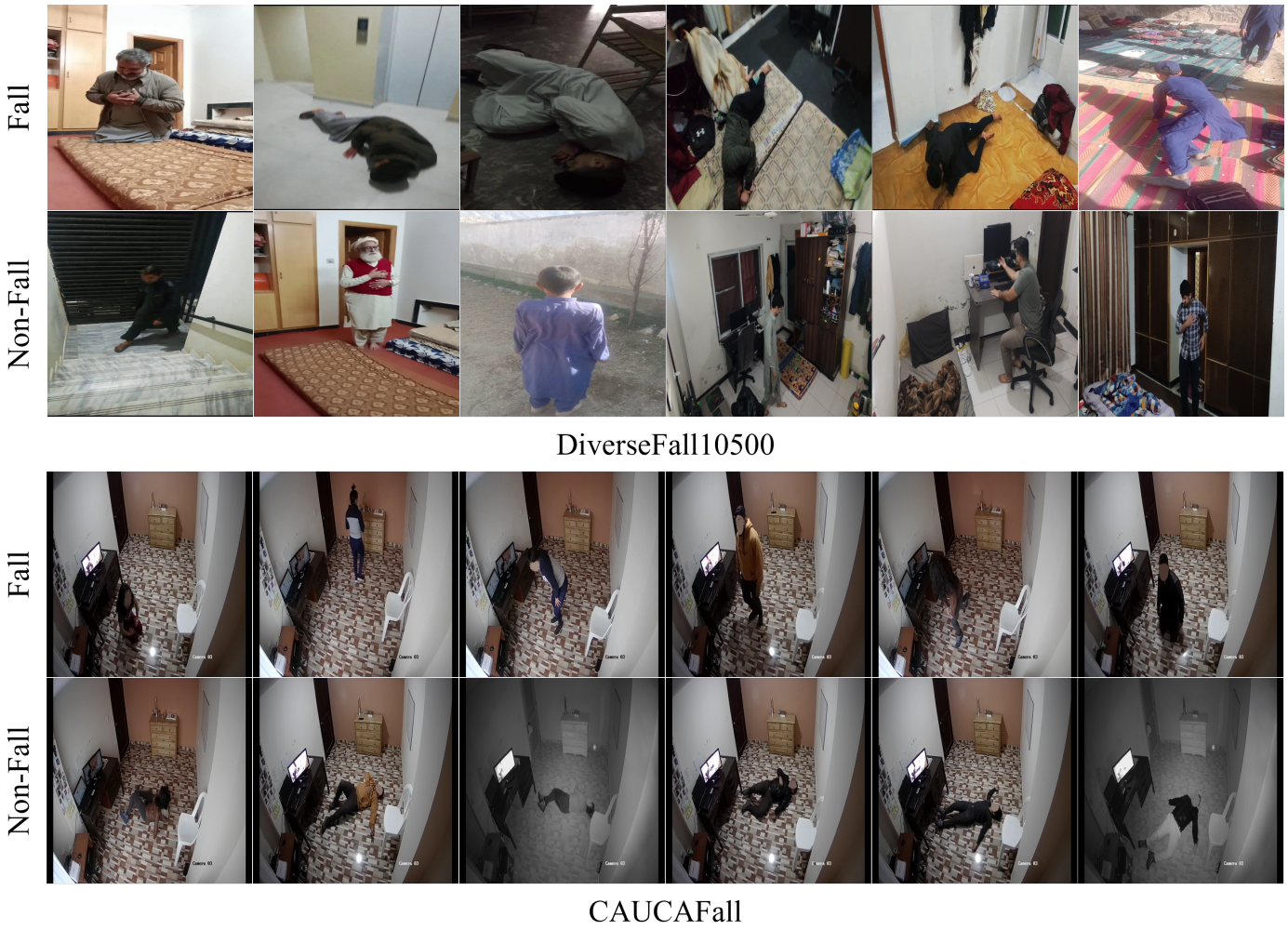

Figure 3. Visual samples for Fall and Non-Fall instances from the DiverseFall and CAUCAFall datasets.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

4.3.2 *Detection Quality Metrics*

We utilized Intersection-over-Union (IoU) based mean Average Precision (mAP) as the primary metric to evaluate detection accuracy:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{13}$$

where N is the number of IoU thresholds and AP_i denotes the Average Precision at the i-th threshold.

4.4 Results Analysis

This section provides a comprehensive analysis of the results obtained from the experiments conducted with the proposed FD network. It includes quantitative evaluations, ablation studies, qualitative analysis, and an assessment of computational complexity, offering insights into the model's performance, effectiveness, and practical feasibility in real-world scenarios.

4.4.1 *Ablation Studies*

We conducted systematic ablation experiments to evaluate the individual contributions of our proposed architectural modifications and quantify their impact on detection performance. These studies provide insights into the efficacy of each attention mechanism and validate our design choices. Tables 2, 3, and 4 present the results of these experiments across multiple performance metrics and computational efficiency indicators.

Incremental Integration of Attention Mechanisms Table 2 illustrates the progressive performance

Table 2. Performance comparison of module integration across DiverseFall and CAUCAFall datasets.

Models	DiverseFall				CAUCAFall			
	mAP	Precision	F1-Score	Recall	mAP	Precision	F1-Score	Recall
YOLOv8n	0.863	0.810	0.822	0.837	0.9925	0.9921	0.9923	0.9935
YOLOv8n + Focus	0.866	0.813	0.827	0.842	0.9925	0.9921	0.9927	0.9937
YOLOv8n + Focus + SE	0.871	0.817	0.832	0.848	0.9926	0.9922	0.9932	0.9941
YOLOv8n + Focus + SE + SA	0.875	0.820	0.838	0.854	0.9927	0.9923	0.9936	0.9944
Proposed network	**0.878**	**0.823**	**0.845**	**0.859**	**0.9927**	**0.9923**	**0.9939**	**0.9946**

Table 3. Ablation study on the effect of different kernel sizes in the SA module across DiverseFall and CAUCAFall datasets.

Kernel Size	Dataset	mAP	F1-Score	Recall
1×1	DiverseFall	0.870	0.832	0.846
	CAUCAFall	0.9926	0.9931	0.9938
3×3	DiverseFall	0.873	0.836	0.851
	CAUCAFall	0.9926	0.9934	0.9941
7×7	DiverseFall	0.875	0.838	0.854
	CAUCAFall	0.9927	0.9936	0.9944

improvements achieved through the sequential integration of different modules into the baseline YOLOv8n architecture. The baseline YOLOv8n model achieved an mAP of 0.863 on the DiverseFall dataset and 0.9925 on the CAUCAFall dataset. The addition of the Focus module provided a modest improvement, increasing the mAP to 0.866 (+0.003) on DiverseFall while maintaining performance on CAUCAFall. When SE blocks were integrated into the early stages (Stages 1–2) of the network, we observed a more significant improvement with mAP increasing to 0.871 (+0.008 from baseline) on DiverseFall and slight gains on CAUCAFall. This confirms our hypothesis that channel attention in early stages helps the model focus on relevant feature channels during the initial feature extraction process.

Further enhancement was achieved by incorporating SA blocks in the later stages (Stages 3–4), resulting in an mAP of 0.875 (+0.012 from baseline) on DiverseFall and continued improvement on CAUCAFall. This validates the effectiveness of spatial attention in later stages, where detecting important spatial regions becomes crucial for accurate fall detection. Finally, our complete proposed architecture, which additionally integrates ECA blocks in the neck section, achieved the best performance with an mAP of 0.878 (+0.015 from baseline) on DiverseFall and consistent improvements across all metrics on both datasets. The F1-score improved from 0.822 to 0.845 (+0.023) on DiverseFall, and from 0.9923 to 0.9939 (+0.0016) on CAUCAFall, demonstrating the balanced enhancement in both precision and recall. These results confirm that our strategy of integrating different attention mechanisms at strategic locations within the network architecture yields complementary benefits, with each module contributing to the overall improvement in performance. The gains were more pronounced on the more challenging DiverseFall dataset, which contains a broader range of fall scenarios and environmental conditions, suggesting that the attention mechanisms enhance the model's ability to distinguish falls across diverse contexts.

Spatial Attention Kernel Size Optimization Table 3 presents an analysis of the effect of different kernel sizes in the Spatial Attention (SA) module. We experimented with three kernel sizes: 1×1, 3×3, and 7×7, evaluating their impact on both datasets. The results demonstrate a clear trend where larger kernel sizes lead to better performance. On the DiverseFall dataset, the mAP increased from 0.870 with a 1×1 kernel to 0.875 with a 7×7 kernel. Similarly, the F1-score improved from 0.832 to 0.838, and recall from 0.846 to 0.854. The CAUCAFall dataset showed similar trends, with more minor absolute improvements due to the already high baseline performance. These findings support our design choice of using a 7×7 kernel in the SA module, as it enables the model to capture broader spatial contexts and relationships between distant parts of the image. This is particularly important for fall detection, where the spatial configuration and relationship between different body parts provide crucial cues for identifying fall events. The larger receptive field enables the model to more accurately capture the distinctive posture changes and spatial patterns associated with falls, thereby contributing to improved detection accuracy.

Table 4. A comparison of the proposed attention-enhanced network with various YOLOv8 versions, considering GFLOPs, parameters, FPS, and model size (MS).

Methods	GFLOPs	Parameters (M)	FPS	MS (MB)	mAP(1)
YOLOv8x	258.5	68.2	83.8	130.5	0.903
YOLOv8l	165.7	43.7	97.4	83.7	0.871
YOLOv8m	79.3	25.9	98.3	49.7	0.899
YOLOv8s	28.8	11.2	130.9	21.5	0.920
YOLOv8n	8.9	3.2	118.0	6.2	0.863
Proposed network	**9.7**	**3.4**	**106.2**	**6.7**	**0.878**

Table 5. Quantitative analysis of our model with different SOTA object detection models on DiverseFall and CAUCAFall datasets.

Models	DiverseFall				CAUCAFall			
	mAP	Precision	F1-Score	Recall	mAP	Precision	F1-Score	Recall
Faster R-CNN	0.831	0.837	0.813	0.809	0.9903	0.9891	0.9902	0.9924
yolov3	0.842	0.848	0.827	0.809	0.9912	0.9904	0.9916	0.9931
yolov4	0.825	0.818	0.801	0.794	0.9896	0.9901	0.9898	0.9913
yolov5	0.870	0.810	0.839	0.852	0.9924	0.9921	0.9935	0.9942
yolov8	0.863	0.810	0.822	0.837	0.9925	0.9921	0.9923	0.9935
Proposed network	**0.878**	**0.823**	**0.845**	**0.859**	**0.9927**	**0.9923**	**0.9939**	**0.9946**

Computational Efficiency Analysis Table 4 presents a comprehensive comparison of our proposed network with different YOLOv8 variants in terms of computational efficiency and accuracy. As shown in the table, our attention-enhanced YOLOv8n model achieves an optimal balance between performance and computational demands. The proposed network requires 9.7 GFLOPs, representing only a 9% increase over the baseline YOLOv8n (8.9 GFLOPs), while significantly outperforming it in terms of mAP (0.878 vs. 0.863). Similarly, the parameter count increased by only 6% (from 3.2M to 3.4M), resulting in a minimal increase in model size from 6.2 MB to 6.7 MB. This efficiency is particularly noteworthy when compared to larger models, such as YOLOv8l, which achieves a comparable mAP (0.871) but requires approximately 17 times more GFLOPs and 13 times more parameters. In terms of inference speed, our model operates at 106.2 FPS, which is only 10% slower than the baseline YOLOv8n (118.0 FPS) but still well within the requirements for real-time applications. This moderate decrease in inference speed is a reasonable trade-off considering the significant gain in detection accuracy. These results validate our approach of strategically integrating lightweight attention mechanisms at key points in the architecture, rather than scaling up the entire model. By targeting specific components of the network with appropriate attention mechanisms, we achieved a performance level comparable to that of significantly larger models, while maintaining the computational efficiency required for deployment on resource-constrained devices. These findings collectively support our architectural design choices and demonstrate the effectiveness of the proposed attention-enhanced YOLOv8 model for fall detection applications.

4.4.2 Quantitative Evaluations

Table 5 compares our proposed attention-enhanced YOLOv8 architecture with state-of-the-art detectors, including Faster R-CNN and YOLOv3-v5 variants,

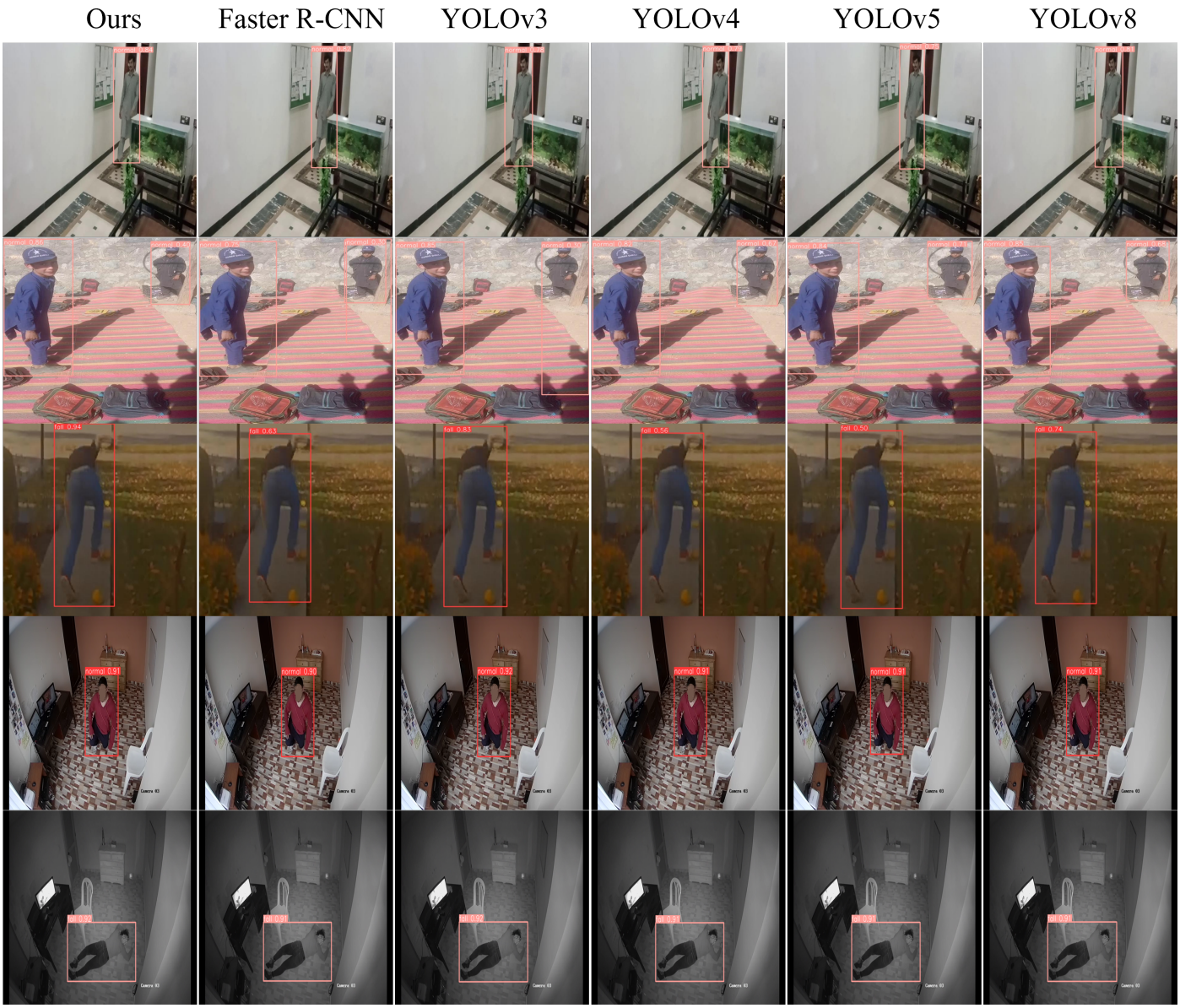

Figure 4. Qualitative comparison of our model with SOTA approaches on DiverseFall and CAUCAFall dataset.

across both fall detection datasets. Our quantitative analysis reveals several key findings:

Performance on DiverseFall Dataset On the more challenging DiverseFall dataset, our proposed network achieves the highest performance across all metrics (mAP: 0.878, precision: 0.823, F1-score: 0.845, recall: 0.859), significantly outperforming the baseline YOLOv8 model with improvements of 1.5% in mAP, 1.3% in precision, 2.3% in F1-score, and 2.2% in recall. YOLOv5 achieves the closest performance with an mAP of 0.870, but our model still surpasses it by 0.8% in mAP and 1.3% in precision. Notably, Faster R-CNN, despite its more complex two-stage architecture, achieves an mAP of only 0.831, which is 4.7% lower than our approach.

Performance on CAUCAFall Dataset On the CAUCAFall dataset, which features more controlled environments, our model maintains its leading position with the highest scores across all metrics (mAP: 0.9927, precision: 0.9923, F1-score: 0.9939, recall: 0.9946). The performance gains over the baseline YOLOv8 are more modest in this dataset, due to the generally high detection accuracy across all approaches; however, our model still demonstrates consistent improvements. Earlier YOLO variants (v3 and v4) and Faster R-CNN demonstrate lower performance, with mAP values 0.15%-0.31% lower than our approach.

Cross-Dataset Analysis The more substantial improvements achieved on the DiverseFall dataset compared to CAUCAFall highlight a key advantage

of our approach: enhanced ability to handle complex and diverse fall scenarios. The integrated attention mechanisms enable the model to adaptively focus on discriminative features regardless of environmental variations, leading to more consistent performance across different scenarios. This is particularly important for real-world fall detection applications, where environmental conditions are often unpredictable and diverse. Additionally, we observe that our model achieves balanced improvements across both precision and recall metrics, indicating enhanced capability in correctly identifying falls and capturing all fall instances. This balance is critical for fall detection systems, where both false positives and false negatives can have significant practical consequences. The quantitative evaluation confirms that our attention-enhanced YOLOv8 architecture effectively improves fall detection performance while incurring minimal computational overhead, demonstrating particular strength in handling diverse and challenging scenarios.

4.4.3 *Qualitative Analysis*

Figure 4 presents a visual comparison between our proposed method and various YOLO variants. Each column contains five different samples. In particular, the first column highlights the robust detection capabilities of our process, often surpassing the baseline YOLO models in terms of precision and clarity. These visual results confirm that our attention-enhanced YOLOv8n architecture not only achieves high mAP scores but also effectively distinguishes between fall and non-fall instances, demonstrating robust generalization and precise localization in diverse environments.

5 Conclusion and Future Work

This work presents a lightweight cascaded feature reweighting approach for fall detection, incorporating complementary attention mechanisms strategically within the YOLOv8 architecture while maintaining computational efficiency. Our key insight lies in recognizing that different network stages have distinct information processing requirements, which benefit from specialized attention types. The architecture's effectiveness stems from a deliberate placement strategy: Squeeze-and-Excitation blocks in the early stages (1–2) selectively emphasize important feature channels during initial feature extraction, capturing essential low-level visual patterns. Meanwhile, Spatial Attention modules with optimized 7×7 kernels in the later stages (3–4) focus on precisely localizing relevant body regions and postural configurations critical for fall identification. Finally, Efficient Channel Attention in the neck section preserves detailed information during feature fusion without dimensionality reduction, maintaining the rich feature representations necessary for accurate classification. This cascaded attention approach creates an information flow where each stage builds upon the refined features of previous stages, progressively focusing computational resources on the most discriminative aspects of the input. Our ablation studies confirm that this specific arrangement outperforms the uniform application of any single attention mechanism, highlighting the complementary nature of the proposed components. In the future, we aim to explore fall detection by combining inputs from multiple sensors, such as visual sensors and wearable sensors, to make the predictions more informative and accurate.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Faryal Zahoor is an employee of BRAINS Institute Peshawar, Peshawar, 25000, Pakistan. The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] *Falls*. (2021, April 26). World Health Organization (WHO). Retrieved from https://www.who.int/news-room/fact-sheets/detail/falls

[2] Ren, L., & Peng, Y. (2019). Research of fall detection and fall prevention technologies: A systematic review. *IEEE Access, 7*, 77702-77722. [CrossRef]

[3] Ma, L., Liu, M., Wang, N., Wang, L., Yang, Y., & Wang, H. (2020). Room-level fall detection based on ultra-wideband (UWB) monostatic radar and convolutional long short-term memory (LSTM). *Sensors, 20*(4), 1105. [CrossRef]

[4] Wang, K., Zhan, G., & Chen, W. (2019). A new approach for IoT-based fall detection system using commodity mmWave sensors. In *Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City* (pp. 197-201). [CrossRef]

[5] Sheng-lan, Z., Yi-fan, Y., Li-fu, G., & Diao, W. (2019, November). Research and design of a fall detection system based on multi-axis sensor. In *Proceedings of the 4th International Conference on Intelligent Information Processing* (pp. 303-309). [CrossRef]

[6] Er, P. V., & Tan, K. K. (2020). Wearable solution for robust fall detection. In *Assistive Technology for the Elderly* (pp. 81-105). Academic Press. [CrossRef]

[7] Charfi, I., Miteran, J., Dubois, J., Atri, M., & Tourki, R. (2012, November). 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* (pp. 218-224). IEEE. [CrossRef]

[8] Mastorakis, G., & Makris, D. (2014). Fall detection system using Kinect's infrared sensor. *Journal of Real-Time Image Processing, 9*(4), 635-646. [CrossRef]

[9] Alam, E., Sufian, A., Dutta, P., & Leo, M. (2022). Vision-based human fall detection systems using deep learning: A review. *Computers in Biology and Medicine, 146*, 105626. [CrossRef]

[10] Zhang, Z., Conly, C., & Athitsos, V. (2014, December). Evaluating depth-based computer vision methods for fall detection under occlusions. In *International symposium on visual computing* (pp. 196-207). Cham: Springer International Publishing. [CrossRef]

[11] Raza, A., Yousaf, M. H., & Velastin, S. A. (2022, June). Human fall detection using yolo: A real-time and ai-on-the-edge perspective. In *2022 12th International Conference on Pattern Recognition Systems* (*ICPRS*) (pp. 1-6). IEEE. [CrossRef]

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

[13] Wang, X., Ellul, J., & Azzopardi, G. (2020). Elderly fall detection systems: A literature survey. *Frontiers in Robotics and AI, 7*, 71. [CrossRef]

[14] Qi, P., Chiaro, D., & Piccialli, F. (2023). FL-FD: Federated learning-based fall detection with multimodal data fusion. *Information Fusion, 99*, 101890. [CrossRef]

[15] Galvão, Y. M., Ferreira, J., Albuquerque, V. A., Barros, P., & Fernandes, B. J. T. (2021). A multimodal approach using deep learning for fall detection. *Expert Systems with Applications, 168*, 114226. [CrossRef]

[16] Lee, E., Kim, J. S., Park, D. K., & Whangbo, T. (2024). YOLO-MR: Meta-Learning-Based Lesion Detection Algorithm for Resolving Data Imbalance. *IEEE Access, 12*, 49762-49771. [CrossRef]

[17] An, J. (2024). Route Positioning System for Campus Shuttle Bus Service Using a Single Camera. *Electronics, 13*(11), 2004. [CrossRef]

[18] Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 6517-6525). [CrossRef]

[19] Killian, L., Maitre, J., Bouchard, K., Lussier, M., Bottari, C., Couture, M., Bier, N., Giroux, S., & Gaboury, S. (2021). Fall prevention and detection in smart homes using monocular cameras and an interactive social robot. In *Proceedings of the Conference on Information Technology for Social Good* (pp. 7-12). [CrossRef]

[20] Zhao, D., Song, T., Gao, J., Li, D., & Niu, Y. (2024). Yolo-fall: A novel convolutional neural network model for fall detection in open spaces. *IEEE Access, 12*, 26137-26149. [CrossRef]

[21] Khekan, A. R., Aghdasi, H. S., & Salehpour, P. (2024). The impact of YOLO Algorithms within fall detection application: A review. *IEEE Access, 13*, 6793-6809. [CrossRef]

[22] Papan, V., & Maheswari, S. (2024, August). Intelligent Fall Detection and Alert System for the Elderly Using Yolov8 and Cloud-Based Analytics. In *2024 5th International Conference on Electronics and Sustainable Communication Systems* (*ICESC*) (pp. 580-588). IEEE. [CrossRef]

[23] Gaya-Morey, F. X., Manresa-Yee, C., & Buades-Rubio, J. M. (2024). Deep learning for computer vision based activity recognition and fall detection of the elderly: a systematic review. *Applied Intelligence, 54*(19), 8982-9007. [CrossRef]

[24] Usman, M. T., Khan, H., Rida, I., & Koo, J. (2025). Lightweight transformer-driven multi-scale trapezoidal attention network for saliency detection. *Engineering Applications of Artificial Intelligence, 155*, 110917. [CrossRef]

[25] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media, 8*(3), 331-368. [CrossRef]

[26] Chen, H., Gu, W., Zhang, Q., Li, X., & Jiang, X. (2024). Integrating attention mechanism and multi-scale feature extraction for fall detection. *Heliyon, 10*(10). [CrossRef]

[27] Kwolek, B., & Kepski, M. (2015). Improving fall detection by the use of depth sensor and accelerometer. *Neurocomputing, 168*, 637-645. [CrossRef]

[28] Yacchirema, D., De Puga, J. S., Palau, C., & Esteve, M. (2018). Fall detection system for elderly people using IoT and big data. *Procedia Computer Science, 130*, 603-610. [CrossRef]

[29] Seredin, O. S., Kopylov, A. V., Huang, S.-C., & Rodionov, D. S. (2019). A skeleton features-based fall detection using Microsoft Kinect v2 with one class-classifier outlier removal. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*, 189-195. [CrossRef]

[30] Chen, L., Li, R., Zhang, H., Tian, L., & Chen, N. (2019). Intelligent fall detection method based on

accelerometer data from a wrist-worn smart watch. *Measurement, 140*, 215-226. [CrossRef]

[31] Chen, T., Ding, Z., & Li, B. (2022). Elderly Fall Detection Based on Improved YOLOv5s Network. *IEEE Access, 10*, 91273-91282. [CrossRef]

[32] Ke, Y., Yao, Y., Xie, Z., Xie, H., Lin, H., & Dong, C. (2023). Empowering Intelligent Home Safety: Indoor Family Fall Detection with YOLOv5. In *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)* (pp. 0942-0949). [CrossRef]

[33] Khan, H., Ullah, I., Shabaz, M., Omer, M. F., Usman, M. T., Guellil, M. S., & Koo, J. (2024). Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset. *Image and Vision Computing, 149*, 105195. [CrossRef]

[34] Lezzar, F., Benmerzoug, D., & Kitouni, I. (2020). Camera-based fall detection system for the elderly with occlusion recognition. *Applied Medical Informatics, 42*(3), 169-179.

Farhan Ali is a graduate student currently pursuing a Master's in Computer Science with a specialization in Data Science at Technische Universität Graz, Austria. He has a strong academic background and hands-on experience in data science, machine learning, and computer vision. He worked as an Associate Software Engineer at OpusAI, where he was involved in building user-centric web applications. In addition, he completed data science internships with Oasis Infobyte and Info AidTech, respectively, gaining valuable experience.

Alexandros Gazis received his diploma in Electronic and Computer Engineering and his MSc in Microelectronics and Computer Systems from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2016 and 2018, respectively. Since 2018, he has been a PhD candidate in the field of computer science at the same university, where he is a member of the "Operating Systems and Middleware for Pervasive Computing and Wireless Sensor Networks" research group. He is also currently pursuing an MBA at Heriot-Watt University since February 2023. Moreover, he is a Teaching Assistant and Lab Demonstrator, supervised by Assistant Professor Eleftheria Katsiri.

Mr. Gazis is a member of the Technical Chamber of Greece and works in the private sector as a Software Engineer for Piraeus Bank S.A., specializing in banking systems. He has published articles on Artificial Intelligence, game engines, web data analytics, remote sensing, and neural networks.

His research focuses on the Internet of Things via wireless sensor networks, cloud computing, and middleware development for pervasive computing. (Email: agazis@ee.duth.gr)

Faryal Zahoor received her BS degree in Computer Science from the University of Agriculture, Peshawar, Pakistan, and MS degree in Computer Science from Islamia College University, Peshawar. Her research interests include computer vision, machine learning, deep learning, medical image processing, and pattern recognition. She is also affiliated with the BRAINS Institute, Peshawar, where she serves as a Lecturer. (Email: faryal.zahoorjan@gmail.com)