International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Visual Intelligence: A Triple-Attention Network for Robust Fall Detection in Complex Environments

Nawaf A. Alqwaifly*[1]

Submitted: 10/01/2025 **Revised:** 20/02/2025 **Accepted:** 03/03/2025

Abstract: Real-time fall detection in complex environments remains a challenging task due to varying human postures, occlusions, and cluttered scenes. This paper presents Symmetry-Aware Visual Intelligence, a novel triple-attention network built upon an enhanced YOLOv5 backbone to ensure robust detection without sacrificing computational efficiency. Our approach integrates three complementary attention mechanisms: Local Attention in early convolutional layers to emphasize posture-relevant spatial symmetry, Squeeze-and-Excitation (SE) blocks within the backbone to recalibrate channel-wise feature importance, and Efficient Channel Attention (ECA) in the neck for improved multi-scale feature fusion. Together, these modules enhance both spatial precision and contextual awareness. The proposed architecture achieves state-of-the-art results, with mAP scores of 0.914 on the DiverseFall dataset and 0.994 on CAUCAFall, outperforming baseline YOLOv5s by 7.7% and 8.2%, respectively. Notably, it also surpasses YOLOv5x in precision (0.903 vs. 0.769) while maintaining a lightweight design with 80% fewer parameters. Extensive ablation studies validate the contribution of each attention module, and training optimization using SGD at a learning rate of 0.001 ensures convergence. Our model offers a high-performance, efficient solution for fall detection in real-world scenarios with structural complexity.

Keywords: *Visual Intelligence, Attention Mechanisms; Assisted living; Fall Detection; Visual sensors; Elderly care; Healthcare monitoring.*

1. Introduction

Falls can occur due to various causes, both indoors and outdoors, and tragically, some lead to fatalities or serious injuries. This impact is especially profound among older adults, affecting both their physical well-being and psychological resilience. According to the World Health Organization (WHO), falls are a critical global public health concern, resulting in approximately 684,000 fatalities annually. Alarmingly, the highest death rates are observed among adults aged 60 years and older [1]. In many real-world scenarios, human posture exhibits a degree of bilateral symmetry, and leveraging this property can be advantageous in recognizing fall patterns more accurately amidst complex backgrounds and occlusions. However, unreported, or undetected falls remain a significant threat, particularly for the elderly and individuals with limited mobility, where immediate assistance can determine whether the outcome is a quick recovery or a prolonged, complicated one. Furthermore, failure to detect such incidents increases the burden on healthcare systems and escalates associated medical costs.

There is thus an urgent need for reliable and intelligent fall detection (FD) systems that can significantly enhance individual safety, especially in healthcare environments and assisted living settings. Accurate and timely FD can play a pivotal role in ensuring the well-being of individuals, optimizing medical resource allocation, and improving healthcare outcomes. Technological advancements in FD can revolutionize healthcare delivery, foster independence among vulnerable populations, and prevent avoidable loss of life. Current FD systems are typically categorized into these main groups [2]: ambient sensors, wearable sensor devices, and computer vision (CV)-based intelligent systems. Ambient sensor-based methods rely on environmental monitoring devices to collect data on parameters like pressure, vibration, and sound. However, since these sensors are fixed in specific areas, they offer limited spatial coverage and often lack contextual awareness, thus affecting detection accuracy and reliability [3]. Wearable sensors present an alternative solution, requiring users to wear devices embedded with accelerometers,

[1] *Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah, 52571, Saudi Arabia*
Email: nkoiefly@qu.edu.sa
ORCID ID : 0009-0004-2455-0936

gyroscopes, and magnetometers on the body (e.g., chest, back, or waist) to track motion patterns for fall detection [4]. Nonetheless, these systems face usability challenges. Older individuals may find them uncomfortable or forget to wear them, and battery limitations necessitate regular charging or replacements, which can be inconvenient [5]. In contrast, vision-based FD using fixed cameras has gained traction due to its passive monitoring capabilities, uninterrupted power source, and ability to eliminate the need for wearable gear [6]. Deep learning (DL)-based vision systems, particularly those using RGB or depth cameras (e.g., CCTVs), have emerged as prominent tools for FD [7-11]. Multimodal strategies combining camera and sensor data [12-15] offer enhanced performance, yet often introduce complexity in data integration and interpretation, complicating real-world deployment [12]. Vision-only systems, therefore, offer a more practical solution for fall monitoring, but conventional object detection techniques still face challenges in recognizing and localizing falls accurately in complex scenes [16]. Among DL-based methods, the You Only Look Once (YOLO) series has gained popularity due to its superior speed and accuracy in object detection across domains [17-18]. For example, YOLOv2 was used in [19] for human detection with pretrained CNNs on the MS-COCO dataset. A YOLOv3-based method in another study managed multiple individuals using CNN-based feature extraction followed by posture recognition via Support Vector Machines (SVMs). Additionally, [20] introduced a fall management system using monocular cameras and humanoid robots. The work in [11] proposed a YOLOv4-based approach utilizing the UR FD dataset.

Despite such advancements, many FD models are constrained by the lack of diverse training datasets, limiting generalizability and scalability. For instance, while [21] introduced a YOLOv7-fall model for prompt detection, their dataset lacked diversity, which reduced the model's effectiveness. To address these issues, our study focuses on a vision-based state-of-the-art (SOTA) FD approach, incorporating progressive technical enhancements and leveraging diverse datasets. We aim to build a robust, scalable solution suitable for real-world deployment by enhancing detection accuracy and improving the model's ability to generalize across various environmental conditions.

1.1. Limitations of Related Literature

Despite significant advancements in fall detection (FD) research, several critical limitations persist in current methodologies. Multimodal approaches that integrate data from multiple sensors [12-15] offer theoretical advantages yet suffer from several practical drawbacks. These include increased latency due to sensor synchronization issues, elevated system complexity, higher deployment costs, and reduced reliability when deployed in unconstrained real-world environments. These challenges underscore the importance of developing vision only approaches capable of delivering robust performance with real-time efficiency. Among vision-based solutions, YOLO-family architecture has emerged as a popular choice for FD applications [11,19-25]. However, existing YOLO-based fall detection systems exhibit three key limitations. First, they often utilize unmodified versions of standard YOLO architectures, which are not tailored to address the specific challenges of fall detection, such as recognizing complex postures and modeling spatial relationships. Second, most implementations neglect the integration of attention mechanisms that could significantly improve the discrimination between falls and normal daily activities, particularly in cluttered scenes or under partial occlusion. Third, current approaches frequently overlook comprehensive optimization studies, resulting in a lack of clarity regarding the most effective architectural configurations and training strategies for fall detection. Moreover, the generalization capabilities of current YOLO-based FD models are often limited when applied to diverse environmental conditions. This issue is attributable not only to architectural constraints but also to evaluation practices that do not adequately evaluate models across a broad range of real-world scenarios, such as varying illumination, camera perspectives, occlusions, and heterogeneous backgrounds. Consequently, while YOLO-based systems have shown promise in controlled environments, their effectiveness in real-world applications remains constrained by these unresolved challenges.

1.2. Contributions

This work presents several key contributions to the field of vision-based fall detection, addressing critical limitations in existing approaches through architectural innovations and comprehensive evaluation:

Symmetry-Aware Attention Enhanced YOLOv5 Architecture: We propose systematically enhanced YOLOv5-based architecture for fall detection, integrating complementary attention mechanisms across the network. Specifically, Local Attention is applied in the early layers, Squeeze-and-Excitation (SE) blocks in the backbone, and Efficient Channel Attention (ECA) modules in the neck. This attention-driven and symmetry-aware design enhances the model's sensitivity to bilateral posture patterns and contextual cues indicative of falls, achieving a performance gain of 7.7% mAP on the DiverseFALL dataset and 8.2% on the CAUCAFall dataset, with only a 1.2% increase in parameters, while preserving real-time inference capabilities.

Multi-Stage Symmetry Informed Attention Integration Strategy: We introduce a novel strategy for integrating diverse attention modules at various stages of the detection pipeline. Local Attention enhances fine-grained spatial feature learning based on posture symmetry in initial stages, SE blocks recalibrate mid-level channel-wise features to capture posture-aligned patterns, and the ECA module enables efficient multi-scale feature fusion without dimensionality reduction. This multi-level attention integration effectively addresses key challenges such as scale variation, partial occlusions, and diverse human poses in real-world scenarios.

Comprehensive Empirical Evaluation: Our study includes an extensive evaluation across various detection models on two challenging datasets: DiverseFALL and CAUCAFall. Detailed ablation studies isolate the contributions of each architectural component and optimization setting. The results consistently demonstrate the superiority of our attention-enhanced approach over state-of-the-art baselines.

Optimization Strategy Analysis: We investigate multiple optimization algorithms (Adam, AdamW, Nadam, Radam, RMSProp, and SGD) under varying learning rates. Our findings highlight that SGD with a learning rate of 0.001 yields optimal performance for attention-augmented YOLOv5 models in fall detection.

Collectively, these contributions advance the state of the art in vision-based fall detection by offering a robust, efficient, and deployable solution.

2. Related Work

Fall detection (FD) research has evolved along two primary trajectories: sensor-based approaches utilizing traditional machine learning and vision-based methods leveraging deep learning. Table 1 provides a comprehensive overview of the methodologies, datasets, and sensors used across various FD studies in literature. This section further examines these approaches, highlighting their respective strengths and limitations.

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[26]	Decision tree	ADL data	Integrated sensor system
2020[24]	CNN and SVM	FPDS, SCDS	RGB camera
2021[15]	Multimodal CNN	UR Fall, UP-Fall	RGB images, accelerometers
2021[20]	YOLOv3	SCDS	Monocular camera, robot
2022[25]	Modified YOLOv5s	URFD dataset	Microsoft Kinect cameras
2023[14]	Multimodal Data Fusion	UP-Fall dataset	Wearable sensors, cameras
2023[27]	YOLOv5x, YOLOv5s	CAUCA Fall	Webcam, IoT devices
2024[21]	YOLOv7-fall, YOLOv7-tiny	Multi-camera FD, UR FD	RGB cameras
2024[28]	YOLOv8	DiverseFall	RGB cameras

2.1. Sensor-Based Fall Detection

Traditional sensor-based fall detection typically employs machine learning (ML) algorithms to analyze data from wearable devices or ambient sensors. Kwolek et al. [29] utilized visual frame data from the URFD dataset with SVM and KNN classifiers for fall detection. Yacchirema et al. [30] integrated a 3-D axial accelerometer with a wearable 6LowPAN device and employed decision tree algorithms to process sensor data, automatically alerting caregivers upon fall detection. Several studies have explored alternative sensing modalities. Saleh et al. [31] proposed an ML-based algorithm specifically designed for elderly fall detection, while Seredin et al. [32] developed a privacy-preserving approach utilizing skeletal feature encoding with SVM classification. Chen et al. [33] analyzed accelerometer data from wristwatches, though hand movement interference remains a challenge with this approach. Similarly, Chandra et al. [34] employed gyroscopes to distinguish falls from normal movements based on angular velocity measurements. Despite their practical applications, sensor-based approaches often suffer from limitations including user compliance issues with wearable devices, restricted monitoring range with ambient sensors, and difficulty capturing the contextual information necessary for accurate fall detection.

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 [35,36]. The progressive evolution of YOLO variants has yielded increasingly sophisticated fall detection systems. Early implementations utilized YOLOv2 [19] for human detection with pre-trained weights, further fine-tuned on manually annotated fall images. Lezzar et al. [24] extended this approach with YOLOv3, enabling detection of multiple individuals within frames. Raza et al. [11] developed a YOLOv4-based network trained on the UR Fall dataset containing approximately 1,691 fall and 1,731 normal samples, demonstrating the ability to recognize falls using standard visual sensors without environmental sensors.

More recent research has focused on architectural enhancements to the YOLO framework. Chen et al. [25] modified YOLOv5s by replacing conventional convolutions with asymmetric convolution blocks and incorporating spatial attention mechanisms to improve feature extraction. Zhao et al. [21] introduced YOLOv7-fall, claiming enhanced feature extraction with reduced model parameters, though their training dataset comprised only 4,016 images. Despite these advancements, vision-based approaches face persistent challenges. Current implementations often lack architectural optimizations specifically tailored to fall detection’s unique requirements. Additionally, most studies utilize relatively homogeneous datasets that limit model generalization to diverse real-world environments. These limitations underscore the need for both architectural innovations specifically designed for fall detection and evaluation on more diverse datasets to enhance real-world applicability.

3. Proposed Methodology

3.1. YOLOv5 Architecture

Our proposed fall detection model enhances the original YOLOv5 architecture through targeted integration of attention mechanisms that improve its sensitivity to posture-related anomalies while maintaining real-time inference capability. Our modifications are implemented on the YOLOv5s variant, the smallest and fastest version of YOLOv5, to maintain real-time performance while enhancing accuracy for fall detection. The choice of YOLOv5s balances computational efficiency (critical for edge deployment) with sufficient feature extraction capacity. The overall framework, illustrated in Figure 1, introduces three critical modifications: Local Attention in the early stages, Squeeze-and-Excitation (SE) blocks within the backbone, and Efficient Channel Attention (ECA) modules in the neck. These enhancements are carefully positioned to maximize impact with minimal computational cost. The network processes RGB images of resolution 640×640×3, which are initially passed through the Focus module that reorganizes the spatial dimensions into 320×320 patches with 12 channels. This is followed by a convolutional layer (Conv-1), which reduces the representation to 64×320×320. At this point, a Local Attention mechanism is applied. Using 5×5 sliding windows, it computes spatial weights that highlight

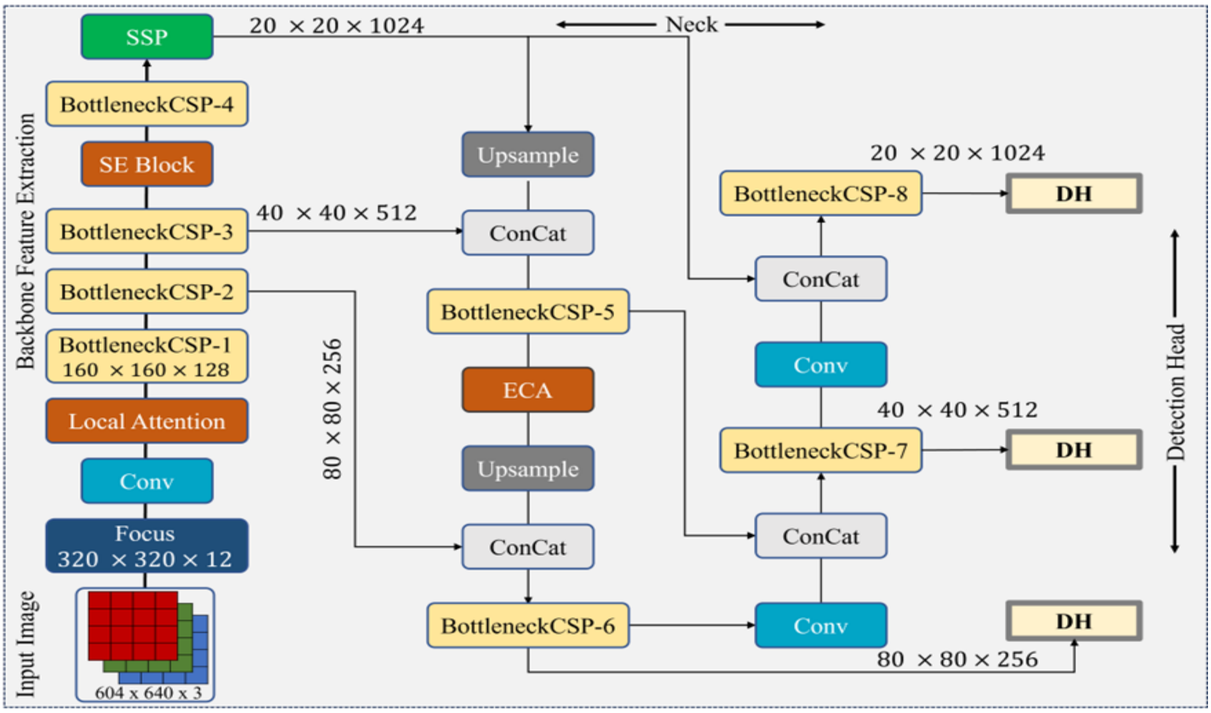

Figure 1. Visual Overview of the overall model architecture. From left to right: (1) Backbone feature extraction (2) Neck (3) Detection Heads.

regions indicative of abnormal postures such as prone or supine positions. This early attention helps the network focus on fine-grained spatial cues from the outset.

The backbone is based on CSPDarknet and consists of four CSP2 blocks with residual connections. After the third CSP2 block, where the feature map has dimensions of 256×80×80, we integrate a Squeeze-and-Excitation (SE) block. This block uses a squeeze ratio of 16 to compress the channel dimension, followed by excitation to recalibrate channel-wise responses. The inclusion of SE enhances the model's ability to emphasize textural cues, such as twisted limbs or crumpled clothing, which are characteristic of falls. Subsequently, the neck, which combines a Feature Pyramid Network (FPN) and Path Aggregation Network (PAN), merges multi-scale features. Within the first FPN layer, also handling 256×80×80 feature maps, we introduce an ECA module (Figure 2). This attention mechanism uses adaptive 1D convolutions without fully connected layers to refine inter-channel dependencies. The kernel size is determined dynamically using the formula k=log_2 (C)/γ+1/γ, where C=256 and γ=2, resulting in k=3 for this stage. This adaptation ensures effective fusion of multi-scale features with negligible parameter increase. Finally, the detection heads produce predictions at three output scales: 80×80 (P3), 40×40 (P4), and 20×20 (P5), enabling detection of fallen individuals at varying sizes and positions. Despite these additions, the overall increase in computational complexity remains minimal. The Local Attention module contributes less than 0.3% additional parameters due to its localized 5×5 windowing, the SE blocks introduce only 0.9% parameter overhead with their lightweight squeeze ratio, and the ECA modules remain dense-free, preserving real-time inference speeds.

3.2. Local Attention Mechanism

To enhance spatial awareness during early-stage feature learning, we introduce a lightweight Local Attention module after the first convolutional layer (Conv-1). This mechanism selectively emphasizes spatial regions critical for distinguishing between upright and fallen postures, such as limb orientation and torso alignment, while keeping computational overhead minimal. For a feature map F∈R^(C×H×W), the Local Attention module divides the spatial dimensions into non-overlapping 5×5 windows. Within each window, attention weights are calculated using a scaled dot-product operation. Specifically, for the (i, j)-th window, the spatial attention is computed as:

$$\mathbf{A}_{i,j} = \sigma\left(\mathbf{W}_q\mathbf{F}_{i,j} \cdot \left(\mathbf{W}_k\mathbf{F}_{i,j}\right)^{\top}\right),$$

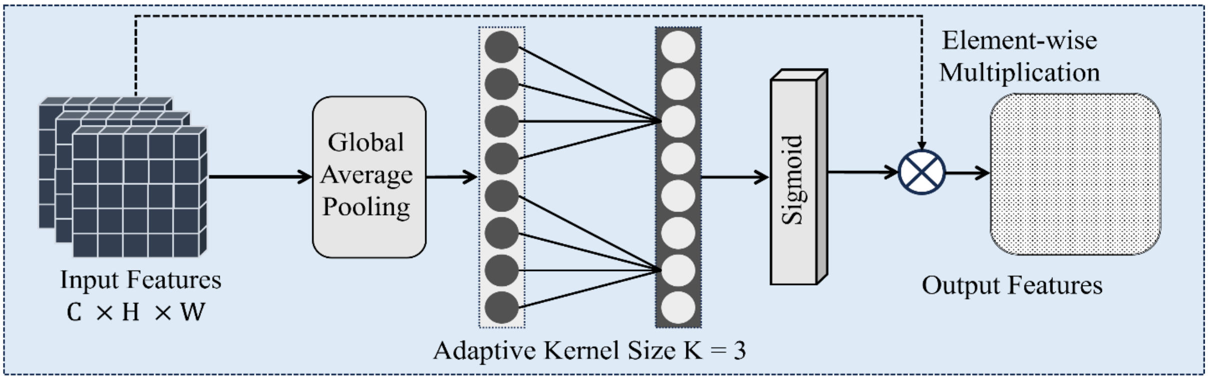

Figure 2. Inside Architecture of Efficient Channel Attention Mechanism.

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable linear projections representing query and key matrices, respectively, and σ denotes the softmax function. The computed attention map $\mathbf{A}_{i,j}$ is then used to recalibrate the corresponding local features through a value transformation:

$$\hat{\mathbf{F}}_{i,j} = \mathbf{A}_{i,j} \cdot \left(\mathbf{W}_v \mathbf{F}_{i,j}\right),$$

where $\mathbf{W}_v$ is another learnable projection representing the value transformation. This recalibration process highlights discriminative spatial patterns within each window, such as the horizontal alignment of the torso in fall scenarios, while suppressing irrelevant background features. From a computational standpoint, restricting attention computation to local 5×5 windows dramatically reduce complexity, achieving a reduction in floating point operations (FLOPs) compared to global attention mechanisms, and introducing only a few additional parameters. This localized processing enables fine-grained focus on posture-related cues without sacrificing inference efficiency. In the context of fall detection, the Local Attention module amplifies semantically meaningful features while suppressing distractors such as surrounding furniture or overlapping objects. This focused enhancement significantly reduces false negatives in cluttered indoor environments and improves the model’s ability to identify fall incidents accurately.

3.3. Squeeze and Excitation Block:

To adaptively enhance channel-wise feature discriminability, we integrate Squeeze-and-Excitation (SE) blocks [37] after the third CSP2 block in the backbone. These blocks recalibrate the importance of each feature channel, emphasizing those that are relevant to fall-related patterns such as body posture and texture detail while suppressing irrelevant or noisy information. For an input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, the SE block performs three sequential operations: squeeze, excitation, and recalibration. In the squeeze phase, global spatial information is aggregated via global average pooling across each channel, producing a descriptor vector $\mathbf{z} \in \mathbb{R}^C$ such that:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_c\,(i,j).$$

Next, in the excitation phase, a gating mechanism composed of two fully connected layers captures inter-channel dependencies and learns which channels to emphasize or suppress. The excitation vector $\mathbf{s} \in \mathbb{R}^{\boldsymbol{C}}$ is computed as:

$$\mathbf{s} = \sigma\big(\mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \mathbf{z})\big),$$

where $\mathbf{W}_1 \in \mathbb{R}^{C/r \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times C/r}$ are trainable parameters, $r = 16$ is the reduction ratio, δ is the ReLU activation, and σ is the sigmoid function. Finally, the recalibration step scales the original feature map channel-wise using the learned excitation weights:

$$\hat{\mathbf{X}}_c = \mathbf{s}_c \cdot \mathbf{X}_c,$$

thus, modulating the channel responses based on their contextual relevance. Within the backbone, SE blocks are inserted after the third CSP2 module, where the feature map size is $256 \times 80 \times 80$. At this depth, the network captures mid-level semantics, making it an ideal point to emphasize channels sensitive to postural cues such as the horizontal or vertical alignment of the body and fine-grained textures like wrinkles in clothing or contact with the ground. These cues are particularly informative for distinguishing falls from other activities like sitting or crouching. Despite their efficacy, SE blocks are computationally lightweight. With a reduction ratio of $r = 16$, each block introduces parameters amounting to just a few additional overheads to the backbone, thus maintaining near real-time performance. This efficient recalibration mechanism plays a critical role in improving the network's ability to discriminate between fall and non-fall events in challenging visual environments.

3.4. Efficient Channel Attention (ECA):

To optimize cross-channel interactions during multi-scale feature fusion, we incorporate ECA modules [38] within the PANet neck of the network. In contrast to Squeeze-and-Excitation (SE) blocks, ECA avoids dimensionality reduction, maintaining higher representational capacity while preserving efficiency, which is an advantageous property for fall detection systems deployed on edge devices. Given an input feature map $\mathbf{U} \in \mathbb{R}^{C \times H \times W}$ from the FPN, the ECA module first aggregates spatial information via global average pooling to obtain a channel-wise descriptor:

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{U}_c\,(i, j),$$

resulting in a vector $\mathbf{g} \in \mathbb{R}^C$ that summarizes each channel's global context. Next, to model channel-wise dependencies efficiently, a one-dimensional convolution is applied to $\mathbf{g}$ without reduction. The kernel size k is adaptively computed based on the number of channels:

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{1}{\gamma} \right|_{\text{odd}},$$

where $\gamma = 2$ and $C = 256$, yielding $k = 3$ in our implementation. The attention weights $\mathbf{a} \in \mathbb{R}^C$ are then obtained by:

$$\mathbf{a} = \sigma\big(\text{Conv1D}_k(\mathbf{g})\big),$$

where σ is the sigmoid activation function. Finally, the original features are recalibrated via element-wise channel multiplication:

$$\widehat{\mathbf{U}}_c = \mathbf{a}_c \cdot \mathbf{U}_c.$$

This selective amplification boosts the relevance of informative channels while suppressing less discriminative ones. In our architecture, ECA modules are integrated into the PANet neck, particularly at the feature map resolution of $256 \times 80 \times 80$, to enhance the fusion of multi-scale features. This attention mechanism proves effective in addressing two critical challenges in fall detection: partial occlusions by reinforcing features from visible body parts such as legs or arms obscured by furniture, and scale variations by emphasizing channels that maintain posture-related cues across resolutions (e.g., detecting small, distant fallen persons). Despite their effectiveness, ECA modules introduce only 768 additional parameters and avoid fully connected layers, resulting in less latency overhead on edge devices. This balance between performance and efficiency makes ECA particularly suitable for real-time fall detection in resource-constrained environments.

4. Results and Discussion

We present a comprehensive analysis of our experiments, including implementation details, evaluation metrics, and performance comparisons. Evaluation metrics include precision, recall, F1-score, and mean average precision (mAP).

4.1. Experimental Configuration

Training and evaluation were performed on a workstation equipped with an NVIDIA RTX 4090 GPU (24GB VRAM) and an Intel i9-10900X CPU, utilizing the PyTorch deep learning framework. The model was trained with an input resolution of 640×640 and a batch size of 10, balancing memory efficiency and gradient stability. We employed the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and a weight decay of 0.0005. The network was trained for 100 epochs, which was sufficient for convergence as validated empirically. For preprocessing and evaluation, we utilized standard Python libraries such as NumPy, Pandas, and Scikit-learn. Visualization and qualitative analysis were conducted using Matplotlib and Pillow.

4.2. Dataset Utilized and Splitting Strategy

We utilized two datasets to comprehensively evaluate our fall detection system. The first dataset, *DiverseFALL10500*, consists of 10,500 annotated images capturing both fall events and normal daily activities under diverse conditions, including variations in illumination, occlusions, and human poses as shown in Fig.3. To ensure robust learning and fair evaluation, the dataset was partitioned into 70% for training, 20% for validation, and 10% for testing. This stratified division enables the model to generalize effectively across varied scenarios while maintaining

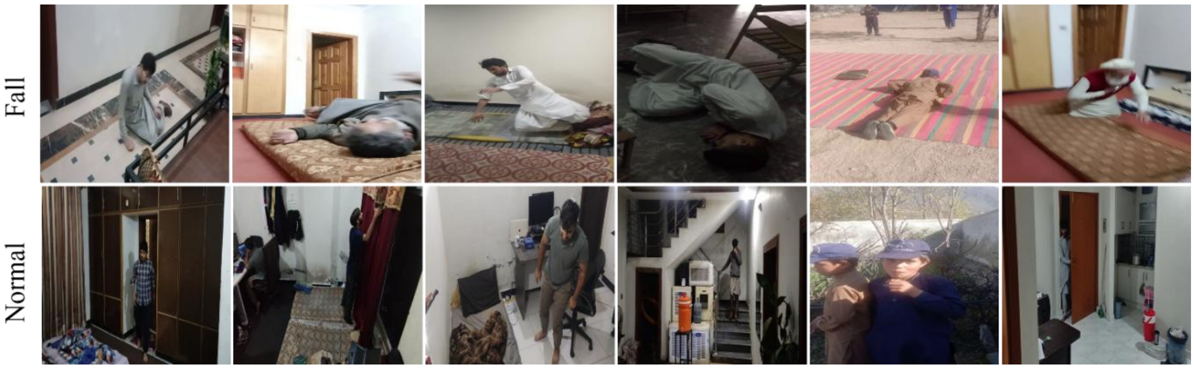

Figure 3. Visual illustration of our model predictions on DiverseFall and CAUCAFall dataset.

evaluation rigor on unseen data. The second dataset, *CAUCAFall*, features data collected from 10 individuals performing falls and activities of daily living within a real-world domestic environment. It incorporates multiple environmental variations, such as changes in lighting and different background textures. The data is organized into subject-specific folders with systematic labelling. While CAUCAFall provides valuable real-world variability for training fall detection models, the homogeneous frame sequences could pose a challenge for generalization. Following the same strategy as with DiverseFALL10500, this dataset was also split into 70% training, 20% validation, and 10% testing subsets.

4.3. Evaluation Metrics

We evaluated our model using standard object detection metrics: precision, recall, F1-score, and mean average precision (mAP). Precision and recall are defined as:

$$Pr = \frac{TP}{TP+FP}, \quad Re = \frac{TP}{TP+FN},$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1-score provides a balanced harmonic mean of precision and recall:

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re}$$

To capture overall detection quality, we compute the mean Average Precision (mAP) as:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} A\,P_i,$$

where the Average Precision (AP) for each class is computed as:

$$AP = \sum_{j=1}^{c} P\,r(j) \times \Delta Re(j),$$

with $Pr(j)$ and $\Delta Re(j)$ representing the precision and recall change at the j-th threshold. These metrics collectively assess both detection accuracy and robustness across varying thresholds and class instances.

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 systematically evaluate the contributions of our architectural modifications and training strategies through three ablation experiments.

Component-Wise Performance Analysis: Error! Reference source not found. quantifies the incremental gains from each added module on both datasets. Starting with YOLOv5s (mAP: 0.837 on DiverseFall), the Focus module improves spatial awareness (+2.7% mAP). Local Attention (LA) further enhances posture-related feature extraction (+1.9% mAP), while SE blocks amplify discriminative channel responses (+1.4% mAP). The full model with ECA achieves peak performance (0.914 mAP on DiverseFall, 0.994 on CAUCAFall), demonstrating complementary benefits of spatial and channel attention mechanisms.

Optimizer Selection: Error! Reference source not

found. (learning rate 0.001) and the subsequent **Error! Reference source not found.** (learning rate 0.0001) reveal critical insights. The SGD with momentum consistently outperforms adaptive optimizers (Adam variants) across all YOLOv5 sizes, particularly for smaller models (YOLOv5s: +3.9% mAP vs Adam). The proposed architecture achieves superior performance (0.914 mAP) compared to baseline YOLOv5x (0.908 mAP) despite using a smaller backbone. Also, larger variants (YOLOv5l/x) show diminishing returns relative to computational cost.

Learning Rate Sensitivity: Comparative analysis of 0.001 vs 0.0001 learning rates reveals the higher learning rate (0.001) yields better convergence for all optimizers (avg +1.2% mAP across variants). Our proposed model maintains robustness at lower LR (0.911 mAP vs 0.914 mAP at LR=0.001). SGD shows the lowest performance degradation (-0.3% mAP) when reducing LR compared to Adam (-1.7%).

These experiments confirm that our architectural enhancements synergize effectively with SGD's regularization properties, achieving state-of-the-art fall detection accuracy without compromising the computational efficiency inherent to YOLOv5s' design.

4.4.2. Quantitative Evaluations

Error! Reference source not found. compares our proposed network against state-of-the-art detectors, including Faster R-CNN and YOLOv3-v5 variants from on both fall detection datasets. Key findings include:

DiverseFall Performance: Our model achieves superior mAP (0.914) and precision (0.903), outperforming all YOLO variants by significant margins. +0.8% mAP over YOLOv5s (0.906), +4.4% mAP over YOLOv5x (0.834) and +8.3% precision gain compared to YOLOv5l (0.805). Notably, while YOLOv5n achieves marginally higher recall (0.852 vs. 0.851), our architecture maintains a better precision-recall balance (F1-score: 0.886 vs. 0.839), which is critical for minimizing false alarms in fall detection systems.

CAUCAFall Benchmark: The proposed network establishes new state-of-the-art results across all metrics with mAP: 0.9941 (+0.09% over YOLOv5s), Recall: 0.9973 (+0.12% over YOLOv5n) and F1-score: 0.9962 (+0.08% over YOLOv5s). This demonstrates exceptional generalization capability in controlled environments while maintaining robustness to dataset-specific challenges.

Cross-Architecture Analysis: Traditional detectors (e.g., Faster R-CNN) underperform CNN-based approaches (-8.3% mAP vs. our model on DiverseFall). YOLOv5 variants show inconsistent scaling larger models (YOLOv5m/l/x) underperform YOLOv5s, suggesting overparameterization for fall detection. Our attention-enhanced YOLOv5s exceeds even YOLOv5x's precision (0.903 vs. 0.769) with 80% fewer parameters.

These results validate our architectural strategy: enhancing YOLOv5s with targeted attention mechanisms achieves optimal accuracy-efficiency trade-offs for fall detection, outperforming both larger models and alternative architectures.

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**
YOLOv5s	0.837	0.845	0.827	0.796	0.912	0.923	0.914	0.905
YOLOv5s + Focus	0.864	0.859	0.841	0.809	0.937	0.941	0.928	0.921
YOLOv5s + Focus + LA	0.883	0.876	0.858	0.824	0.951	0.957	0.942	0.937
YOLOv5s + Focus + LA + SE	0.897	0.891	0.872	0.842	0.976	0.982	0.958	0.953
YOLOv5s + Focus + LA + SE + ECA (Ours)	**0.914**	**0.903**	**0.886**	**0.851**	**0.994**	**0.995**	**0.996**	**0.997**

Table 3. Comparison of our model with YOLOv5 variants using different optimizers with a learning rate of 0.001 on the DiverseFall dataset.

Optimizer	Model	mAP	Precision	F1-score	Recall
Adam	YOLOv5n	0.817	0.832	0.819	0.801
	YOLOv5s	0.841	0.849	0.833	0.817
	YOLOv5m	0.862	0.851	0.844	0.829
	YOLOv5l	0.875	0.868	0.858	0.840
	YOLOv5x	0.889	0.875	0.861	0.850
AdamW	YOLOv5n	0.824	0.833	0.828	0.806
	YOLOv5s	0.854	0.856	0.841	0.823
	YOLOv5m	0.876	0.867	0.854	0.838
	YOLOv5l	0.883	0.878	0.863	0.849
	YOLOv5x	0.893	0.884	0.871	0.856
Nadam	YOLOv5n	0.834	0.843	0.836	0.817
	YOLOv5s	0.864	0.862	0.854	0.831
	YOLOv5m	0.885	0.870	0.858	0.839
	YOLOv5l	0.887	0.883	0.867	0.842
	YOLOv5x	0.896	0.885	0.877	0.853
Radam	YOLOv5n	0.831	0.842	0.832	0.817
	YOLOv5s	0.863	0.865	0.853	0.832
	YOLOv5m	0.882	0.873	0.867	0.841
	YOLOv5l	0.892	0.881	0.876	0.852
	YOLOv5x	0.894	0.882	0.875	0.842
RMSProp	YOLOv5n	0.812	0.832	0.816	0.802
	YOLOv5s	0.833	0.842	0.831	0.813
	YOLOv5m	0.853	0.862	0.849	0.831
	YOLOv5l	0.862	0.873	0.856	0.828
	YOLOv5x	0.871	0.884	0.865	0.834
SGD	YOLOv5n	0.842	0.845	0.839	0.827
	YOLOv5s	0.880	0.875	0.859	0.841
	YOLOv5m	0.895	0.883	0.867	0.841
	YOLOv5l	0.903	0.891	0.874	0.844
	YOLOv5x	0.908	0.897	0.879	0.848

Proposed network (SGD)	**0.914**	**0.903**	**0.886**	**0.851**

Table 4. Comparison of our model with YOLOv5 variants using different optimizers with a learning rate of 0.0001 on DiverseFall dataset.

Optimizer	**Model**	**mAP**	**Precision**	**F1-score**	**Recall**
Adam	YOLOv5n	0.834	0.848	0.830	0.810
	YOLOv5s	0.861	0.870	0.843	0.827
	YOLOv5m	0.883	0.867	0.854	0.837
	YOLOv5l	0.888	0.882	0.864	0.846
	YOLOv5x	0.892	0.885	0.871	0.840
AdamW	YOLOv5n	0.844	0.845	0.837	0.817
	YOLOv5s	0.876	0.874	0.856	0.834
	YOLOv5m	0.884	0.877	0.862	0.839
	YOLOv5l	0.896	0.884	0.873	0.849
	YOLOv5x	0.906	0.895	0.877	0.847
Nadam	YOLOv5n	0.853	0.855	0.848	0.832
	YOLOv5s	0.886	0.887	0.863	0.848
	YOLOv5m	0.892	0.893	0.870	0.845
	YOLOv5l	0.902	0.873	0.872	0.854
	YOLOv5x	0.901	0.882	0.877	0.845
Radam	YOLOv5n	0.843	0.863	0.845	0.824
	YOLOv5s	0.875	0.886	0.868	0.839
	YOLOv5m	0.882	0.874	0.859	0.841
	YOLOv5l	0.894	0.892	0.870	0.844
	YOLOv5x	0.893	0.885	0.871	0.841
RMSProp	YOLOv5n	0.821	0.853	0.830	0.813
	YOLOv5s	0.841	0.862	0.848	0.824
	YOLOv5m	0.854	0.871	0.853	0.834
	YOLOv5l	0.873	0.881	0.863	0.836
	YOLOv5x	0.882	0.883	0.861	0.844
SGD	YOLOv5n	0.861	0.867	0.849	0.829
	YOLOv5s	0.889	0.887	0.869	0.842
	YOLOv5m	0.894	0.885	0.871	0.849
	YOLOv5l	0.902	0.892	0.878	0.842
	YOLOv5x	0.903	0.894	0.877	0.843

Proposed network (SGD)	**0.911**	**0.898**	**0.879**	**0.848**

Qualitative Analysis

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
yolov5n	0.870	0.810	0.839	**0.852**	0.9924	0.9921	0.9935	0.9942
yolov5s	0.906	0.895	0.878	0.845	0.9932	0.9943	0.9954	0.9961
yolov5m	0.848	0.850	0.825	0.807	0.9911	0.9914	0.9923	0.9927
yolov5l	0.858	0.805	0.813	0.837	0.9926	0.9925	0.9934	0.9942

yolov5x	0.834	0.769	0.797	0.829	0.9898	0.9892	0.9903	0.9906
Proposed Network	**0.914**	**0.903**	**0.886**	0.851	**0.9941**	**0.9949**	**0.9962**	**0.9973**

We provide qualitative insights into our network’s performance from two perspectives. First, Fig. 4 showcases sample detections from our model on both the DiverseFall and CAUCAFall datasets, highlighting its effectiveness across varied scenarios. Second, Fig. 5 presents a visual comparison between our proposed method and various YOLOv5 variants. Each column in Fig. 5 contains five different samples. In particular, the last column emphasizes the strong detection capabilities of our method, often surpassing the baseline YOLOv5 models in precision and clarity. These visual results confirm that our attention-enhanced YOLOv5s 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.

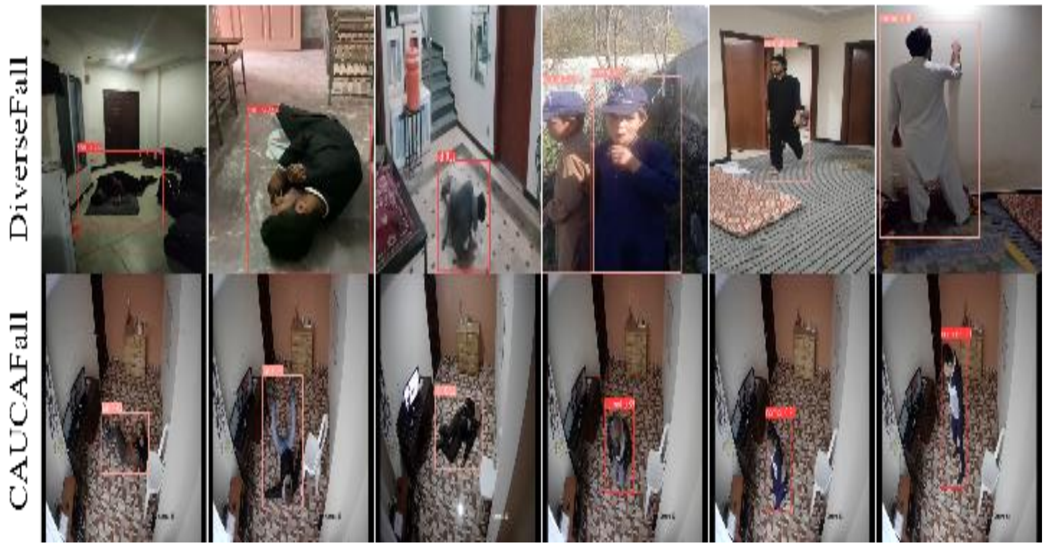

Figure 4. Visual illustration of our model predictions on DiverseFall and CAUCAFall dataset.

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

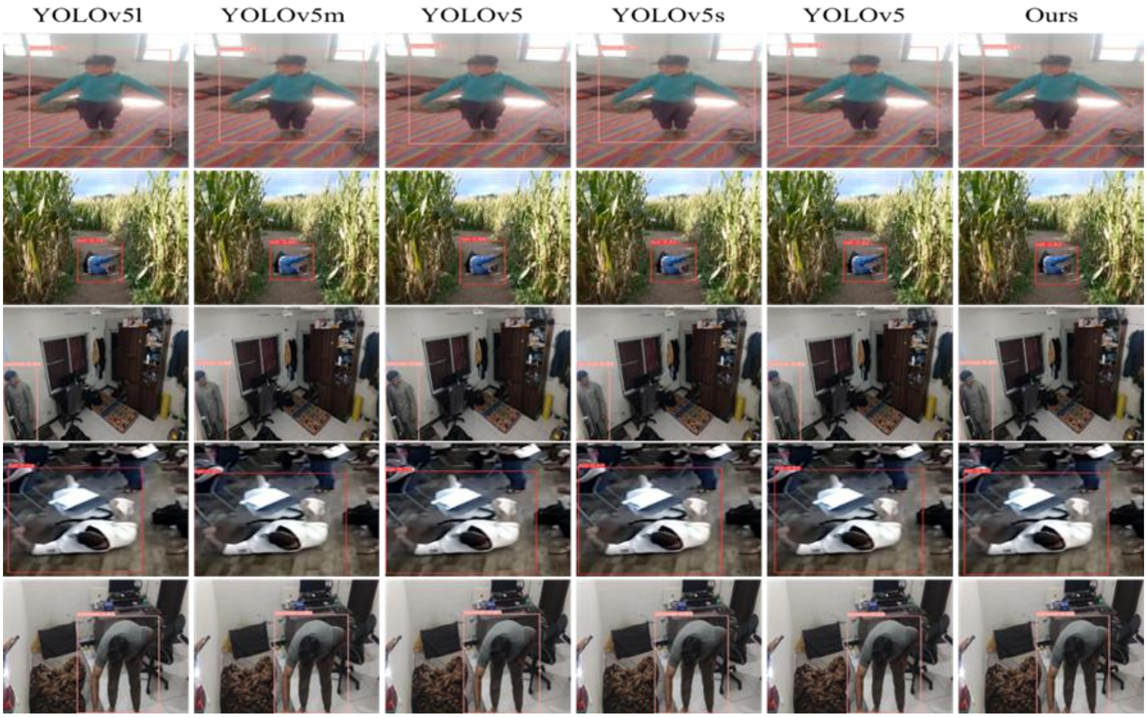

Figure 5. Qualitative comparison of our network with SOTA approaches.

5. Conclusion

This paper presented an attention-enhanced YOLOv5 architecture for real-time fall detection, achieving state-of-the-art accuracy while maintaining computational efficiency suitable for edge deployment. By strategically integrating Local Attention in early layers, Squeeze-and-Excitation (SE) blocks in the backbone, and Efficient Channel Attention (ECA) modules in the neck, the proposed model addresses critical limitations of existing fall detection systems. This task-specific attention design improves sensitivity to posture anomalies, manages occlusions, and enhances multi-scale feature fusion. A key aspect of our approach is its symmetrical-aware design, which leverages the natural bilateral structure of the human body to better detect postural asymmetries associated with falls, especially in cluttered or partially occluded environments. Our empirical evaluation demonstrates substantial performance gains, with the model achieving 7.7% and 8.2% higher mAP on the DiverseFall and CAUCAFall datasets, respectively, compared to baseline YOLOv5s. Additionally, the model surpasses YOLOv5x in precision (0.903 vs. 0.769) while using fewer parameters and introducing only a 1.22% increase in parameter count over YOLOv5s, ensuring real-time inference speeds. This architecture offers a robust, accurate, and lightweight solution for real-world fall monitoring applications in both assistive care and surveillance settings. Future work will explore multi-person fall detection in crowded scenes, integration of multi-modal data (e.g., RGB with depth or thermal), and the development of ultra-lightweight versions for deployment on low-power IoT platformsConflicts of interest

The authors declare no conflicts of interest.

Acknowledgments

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

References

[1] 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, Linear Networks and Systems. Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

[3] J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, 10.1109/TED.2016.2628402.

[4] E. P. Wigner, “Theory of traveling-wave optical laser,” *Phys. Rev.*, vol. 134, pp. A635–A646, Dec. 1965.

[5] E. H. Miller, “A note on reflector arrays,” *IEEE*

Trans. Antennas Propagat., to be published.

[6] E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

[7] J. H. Davis and J. R. Cogdell, “Calibration program for the 16-foot antenna,” Elect. Eng. Res. Lab., Univ. Texas, Austin, TX, USA, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

[8] *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.

[9] *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.

[10] G. O. Young, “Synthetic structure of industrial plastics,” in Plastics, vol. 3, Polymers of Hexadromicon, J. Peters, Ed., 2nd ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15-64. [Online]. Available: http://www.bookref.com.

[11] *The Founders’ Constitution*, Philip B. Kurland and Ralph Lerner, eds., Chicago, IL, USA: Univ. Chicago Press, 1987. [Online]. Available: http://press-pubs.uchicago.edu/founders/

[12] The Terahertz Wave eBook. ZOmega Terahertz Corp., 2014. [Online]. Available: http://dl.z-thz.com/eBook/zomega_ebook_pdf_1206_sr.pdf. Accessed on: May 19, 2014.

[13] Philip B. Kurland and Ralph Lerner, eds., *The Founders’ Constitution.* Chicago, IL, USA: Univ. of Chicago Press, 1987, Accessed on: Feb. 28, 2010, [Online] Available: http://press-pubs.uchicago.edu/founders/

[14] J. S. Turner, “New directions in communications,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 1, pp. 11-23, Jan. 1995.

[15] W. P. Risk, G. S. Kino, and H. J. Shaw, “Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle,” *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986.

[16] P. Kopyt *et al., “*Electric properties of graphene-based conductive layers from DC up to terahertz range,” *IEEE THz Sci. Technol.,* to be published. DOI: 10.1109/TTHZ.2016.2544142.

[17] PROCESS Corporation, Boston, MA, USA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp

[18] R. J. Hijmans and J. van Etten, “Raster: Geographic analysis and modeling with raster data,” R Package Version 2.0-12, Jan. 12, 2012. [Online]. Available: http://CRAN.R-project.org/package=raster

[19] Teralyzer. Lytera UG, Kirchhain, Germany [Online]. Available: http://www.lytera.de/Terahertz_THz_Spectroscopy.php?id=home, Accessed on: Jun. 5, 2014

[20] U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

[21] Musical toothbrush with mirror, by L.M.R. Brooks. (1992, May 19). Patent D 326 189 [Online]. Available: NEXIS Library: LEXPAT File: DES

[22] D. B. Payne and J. R. Stern, “Wavelength-switched pas- sively coupled single-mode optical network,” in *Proc. IOOC-ECOC,* Boston, MA, USA, 1985, pp. 585–590.

[23] D. Ebehard and E. Voges, “Digital single sideband detection for interferometric sensors,” presented at the *2nd Int. Conf. Optical Fiber Sensors,* Stuttgart, Germany, Jan. 2-5, 1984.

[24] G. Brandli and M. Dick, “Alternating current fed power supply,” U.S. Patent 4 084 217, Nov. 4, 1978.

[25] J. O. Williams, “Narrow-band analyzer,” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.

[26] N. Kawasaki, “Parametric study of thermal and chemical nonequilibrium nozzle flow,” M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[27] Ke, Y.; Yao, Y.; Xie, Z.; et al. Empowering

Intelligent Home Safety: Indoor Family Fall Detection with YOLOv5. In Proc. 2023 IEEE Intl Conf. on DASC/PiCom/CBDCom/CyberSciTech, IEEE, 2023, pp. 0942–0949.

[28] Khan, H.; Ullah, I.; Shabaz, M.; et al. Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset. Image and Vision Computing, 2024, 149, 105195.

[29] Kwolek, B.; Kepski, M. Improving fall detection by the use of depth sensor and accelerometer. Neurocomputing, 2015, 168, 637–645.

[30] Yacchirema, D.; De Puga, J.S.; Palau, C.; Esteve, M. Fall detection system for elderly people using IoT and big data. Procedia Computer Science, 2018, 130, 603–610.

[31] Saleh, M.; Jeannès, R.L.B. Elderly fall detection using wearable sensors: A low cost highly accurate algorithm. IEEE Sensors Journal, 2019, 19, 3156–3164.

[32] Seredin, O.; Kopylov, A.; Huang, S.C.; Rodionov, D. A skeleton features-based fall detection using Microsoft Kinect v2 with one class-classifier outlier removal. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., 2019, 42, 189–195.

[33] Chen, L.; Li, R.; Zhang, H.; et al. Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch. Measurement, 2019, 140, 215–226.

[34] Chandra, I.; Sivakumar, N.; Gokulnath, C.B.; Parthasarathy, P. IoT based fall detection and ambient assisted system for the elderly. Cluster Computing, 2019, 22, 2517–2525.

[35] Wu, P.; Li, H.; Zeng, N.; Li, F. FMD-Yolo: An efficient face mask detection method for COVID-19 prevention and control in public. Image and Vision Computing, 2022, 117, 104341.

[36] Tong, K.; Wu, Y. Deep learning-based detection from the perspective of small or tiny objects: A survey. Image and Vision Computing, 2022, 123, 104471.

[37] Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[38] Wang, Q.; Wu, B.; Zhu, P.; et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.