Image and Vision Computing 149 (2024) 105195

Contents lists available at ScienceDirect

Image and Vision Computing

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

Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset

Habib Khan [a], Inam Ullah [b], Mohammad Shabaz [c], Muhammad Faizan Omer [d], Muhammad Talha Usman [a], Mohammed Seghir Guellil [e], JaKeoung Koo [a,*]

[a] School of Computing, Gachon University, 1342 Seongnam-daero, Seongnam-si 13120, South Korea
[b] Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea
[c] Model Institute of Engineering and Technology, Jammu, JK 181122, India
[d] Department of Computer Science, Sir Syed CASE Institute Of Technology, Islamabad 44000, Pakistan
[e] Faculty of Economics, Business and Management Sciences, MCLDL Laboratory, University of Masca, 38489, Algeria

ARTICLE INFO

Keywords:
Assisted living
Fall detection
Computer vision
Data benchmarking
Visual intelligence
Safety monitoring

ABSTRACT

Falls pose a significant risk to elderly people, patients with diseases such as neurological disorders, cardiovascular diseases, and disabled children. This highlights the need for real-time intelligent fall detection (FD) systems for quick relief leading to assisted living. The existing attempts are often based on multimodal approaches which are computationally expensive due to multi-sensor integration. The computer vision (CV) based era for FD needs the deployment of state-of-the-art (SOTA) networks with progressive enhancements to grasp falls effectively. However, CV-based systems often lack the ability to operate efficiently in real-time and the attempts for visual intelligence are usually not integrated at feasible stages of the networks. More importantly, the lack of large-scale well-annotated benchmark datasets limits the ability of FD in challenging and complex environments. To bridge the research gaps, we proposed an enhanced version of YOLOV8 for FD. Our research presents significant contributions by addressing these limitations through three key contributions. Initially, a comprehensive large-scale dataset is introduced which comprises approximately 10,500 image samples with corresponding annotations. The dataset encompasses diverse environmental conditions and scenarios, facilitating the generalization ability for the models. Then, progressive enhancements to the YOLOV8S model are proposed, integrating a focus module in the backbone to optimize feature extraction. Moreover, the convolutional block attention modules (CBAMs) are integrated at the feasible stages of the network to improve spatial and channel contexts for more accurate detection, especially in complex scenes. Finally, an extensive empirical evaluation showcases the superiority of the proposed network over 13 SOTA techniques, substantiated by meticulous benchmarking and qualitative validation across varied environments. The empirical findings and analysis of multiple factors such as model performance, size, and processing time prove that the suggested network displays impressive results. Datasets with annotations, results, and the ways of progressive modifications in the code will be available to the research community at the link https://github.com/habib1402/Fall-Detection-DiverseFall10500

1. Introduction

Falls occur due to various causes which may happen either indoors or outdoors. Tragically, some falls end with deaths or serious injuries. These implications are especially notable among older adults regarding their physical health and mental resilience. According to the World Health Organization (W.H.O), Fall is a critical global public health concern, with approximately 684,000 fatal falls occurring yearly. Alarmingly, death rates are highest among adults aged 60 years and older [1]. Unreported falls are a serious concern for the elderly and those with limited mobility. While immediate assistance may mean the difference between a quick or prolonged and painful recovery. In addition, undetected falls put pressure on healthcare resources and lead to more expensive medical costs. Therefore, there is an urgent demand for reliable and intelligent fall detection (FD) system as it can profoundly influence the overall well-being of individuals, including elders. Precise and timely FD can significantly assist the safety of human lives, particularly in healthcare facilities leading to assisted living. Hence, it is

* Corresponding author.
E-mail addresses: habibkhan@ieee.org (H. Khan), jakeoung@gachon.ac.kr (J. Koo).

https://doi.org/10.1016/j.imavis.2024.105195
Received 12 April 2024; Received in revised form 16 July 2024; Accepted 22 July 2024
Available online 26 July 2024
0262-8856/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

crucial to develop robust and effective FD systems to enhance the standard of care, optimize resource allocation, and improve healthcare results. Advances in FD technologies can revolutionize healthcare operations, promote more independence among vulnerable groups, and ultimately save loss of lives.

The current FD systems can be categorized into three main groups [2]: ambient sensors, wearable sensor devices, and computer vision (CV) based intelligent systems. Environmental sensor-based techniques involve installing monitoring equipment in the living spaces of elderly adults to gather data on pressure, vibration, and sound to identify potential falls. However, these sensors typically installed in specific areas of a living space, may provide limited coverage, as they are confined to detecting falls within their designated range. Moreover, these systems often lack the ability to capture contextual information surrounding falls thus limiting the accuracy and reliability of detection [3]. Wearable sensor devices offer an alternative, which requires users to affix gadgets with magnetometers, gyroscopes, and accelerometers to their back, chest, or waist. The sensor data is collected and analyzed to identify falls in the elderly by tracking their movements [4]. However, these devices require individuals to continuously wear specific gadgets, which may not be practical or acceptable for all users, particularly older adults. Furthermore, limited battery life necessitates frequent recharging or battery replacements, which can be inconvenient for users [5]. On the other hand, visual intelligence-based FD has become a central area of research. In these systems, visual sensor data are analyzed to detect sudden falls in real-time surveillance. Fixed cameras provide an uninterrupted power source for live surveillance, reducing the necessity of using wearable equipment [6].

Among DL-based techniques, vision-driven utilizing regular or depth camera data, such as CCTVs [7–11], constitute the primary strategies for FD. Additionally, multimodal strategies combine data from both sensors and cameras [12–15]. These approaches may offer better performance by utilizing multiple sensors [15], which can lead to increased complexity in data processing and interpretation [12]. Consequently, these approaches encounter notable deployment challenges in real-world applications. Conversely, vision-based systems are frequently employed for passive area monitoring and fallen person detection. While various conventional approaches have been proposed for real-time object detection, these methods often struggle to accurately identify and locate objects in complex environments [16]. The You Only Look Once (YOLO) algorithms have gained prominence for superior performance in different domains [17,18]. For instance, research [19] utilized the YOLOv2 model to detect humans, leveraging pre-trained Convolutional Neural Networks (CNNs) with the MS-COCO dataset. Another study [20] proposed a YOLOV3-based method for FD, considering individuals in frames, and the model was assessed using two datasets. Furthermore, an assisted system introduced in the study [21] utilizes monocular cameras and humanoid robots to manage fall risk. Additionally, the work [11] suggested a YOLOV4-based network and they employed the UR FD dataset. However, despite advancements, the absence of diverse samples are observed for training which limits performance and scalability. Most recently, [22] introduced a YOLOv7-fall model aimed at prompt FD, yet their dataset need the inclusion of more diversity. As such, we prioritize the vision-based state-of-the-art (SOTA) approach with progressive technical modifications and diverse data in our research, recognizing its potential for efficient and effective real-world deployment.

1.1. Limitations of the related literature

It is observable from the literature that multimodal techniques [12–15] have the potential to improve performance by integrating multiple sensors. However, they often demonstrate delays, impracticality, and poor usability due to data integration from multiple sensors. This presents significant difficulties in their implementation and diminishes their appropriateness for real-world applications and practical use. Thus, research on the SOTA networks capable of detecting falls with feasible real-time processing is the need of the day in the FD domain. Moreover, the existing FD lightweight networks especially some from the YOLO family [11,19–25] are not well trained on highly diverse data as the samples are often recorded in homogenous environments which hinders models from good generalization and learning in uncertain environments. These models usually face difficulty in adjusting to variational environments in real-time processing. Furthermore, it is observable that researchers proposed a wide range of networks in the SOTA YOLO family from YOLOV2 to V7 for FD [11,19,21-22,20,25]. However, an advanced YOLO version with progressive modifications to detect challenging falls and to offer improved speed and accuracy over YOLOv2-v7 is imperative in the domain. Additionally, the previous FD datasets often face limitations, including minimal diversity in samples, a limited visual representation of various age groups, insufficient coverage of environmental conditions, a scarcity of high illumination variations, insufficient consideration of diverse lighting conditions, and a constrained diversity in attire representation. These limitations hinder the generalizability of FD models to diverse populations, impacting their real-world applicability. Therefore, the development of a large-scale benchmark dataset with the inclusion of more diversities is also needed for the research community in this domain. Moreover, current SOTA FD methods are still under better development and need improvement in proposing diverse data with progressive modifications in the SOTA network to achieve better accuracy and reduce false alarms.

1.2. Offered contributions

- In response to the highlighted limitations of lacking a diverse dataset in the FD domain, we developed a large-scale challenging FD dataset that includes almost 10,500 images with corresponding annotations. The proposed dataset covered a wide range of environments, different lighting conditions, variations in fall angles, recording samples from different ages of people, incorporating illumination variations, and samples from indoor, and outdoor scenes. The dataset includes diverse scenarios to ensure the adaptability of training FD systems in various environments for real-time processing.
- We optimized the YOLOV8S object detection model for FD by integrating a focus module in the backbone, and convolutional block attention modules (CBAMs)at multiple stages. The enhancement of replacing the convolution with the focus module allows the network to process more spatial information in the early layers without increasing computational costs. Moreover, the integration of CBAMs at strategic stages significantly enhances the ability to focus on the most relevant parts of the features. This multi-stage integration of attention mechanisms improve the feature maps of different levels ultimately boosting performance to detect falls across various scales and complexities.
- We conducted a thorough empirical study using thirteen different detection methods on two datasets. Our suggested network showcased both robustness and higher performance in comparison to SOTA techniques, as evidenced by extensive evaluation and thorough ablation assessments conducted on both established benchmark and proposed datasets. The results are visually validated by a comprehensive qualitative analysis, which confirmed the effectiveness of the attempts for progressive improvements.

1.3. Study outline

The remainder of this work is structured as follows. Section 2 contains a detailed analysis of related work which contains two subcategories includingML-assisted sensors-based approaches and DL-assisted visual intelligence-driven methods. In Section 3, a comprehensive explanation of our network is provided from required technical perspective. Section 4 presents the discussion on the proposed dataset, empirical results, comparative analysis, and discussion. Section 5 concludes the paper with a summary of the findings, their implications, and

potential avenues for future research.

2. Related work

The relevant research in the FD domain can be largely classified into two primary groups of approaches: ML-based approaches, and DL-driven visual intelligence-driven techniques. ML methods are often subject to statistical ways and the method of feature engineering to recognize patterns that are relevant to falls [26]. On the other hand, DL-driven visual intelligence techniques use complex neural network architectures to retrieve complex features from visual samples, allowing for more precise FD. These approaches have shown better results in recent years, demonstrating considerable potential for improving the overall FD systems [22,25]. Table 1 offers a thorough summary of the FD methods employed in different research studies. The demonstrated approaches include a diverse set of methods, including SVM classification algorithms, depth-based analysis, computer vision techniques, machine learning algorithms, multimodal fusion approaches, and deep neural networks-based SOTA approaches such asYOLO variations. These algorithms have been assessed on various datasets using a range of sensors including single cameras, Kinect's infrared sensors, depth cameras, wearable sensors, accelerometers, and IoT devices.

Table 1
Comparative analysis of fall detection techniques in various studies of the related literature. The PD defines that the dataset is the proposed in the paper.

Year and Reference	Technique	Dataset	Senors
2012 [7]	SVM classifier	Realistic dataset	Single camera
2014 [8]	Depth-based 3D bounding box analysis	PD	Kinect's infrared sensor
2014 [10]	Depth-Based Computer Vision Method	Occluded FD dataset	Depth camera
2018 [37]	YOLO v2, UAV	MS COCO, SCDS	RGB
2019 [38]	Multimodal impact features	Le2i FD database, SEIN fall database	Monocular video cameras
2019 [31]	SVM-based fall detection algorithm	SisFall	Wearable sensor
2019 [4]	Body feature estimation algorithm	UP-FD dataset, SisFall: A Fall and Movement Dataset.	Commodity mmWave sensors
2019 [5]	Body posture angle, SVM	Real-time data	MPU6500 sensor
2019 [35]	Enhanced threshold-based	650 test activities	Smartphone built-in accelerometers
2019 [34]	Multisensor fusion-based, Logistic regression	Real-time accelerometer data	Tri-axis accelerometer, gyroscope
2019 [12]	Multimodal Approach	UP-FD Dataset	Wearable sensors, Ambient sensors
2020 [39]	Region based CNN, Transfer learning	Le2i FDD	RGB
2020 [36]	Decision tree	ADL data	Integrated sensor system
2020 [20]	CNN and SVM	FPDS, SCDS	RGB
2021 [15]	Multimodal CNN	UR Fall, UP-Fall	RGB images, accelerometers
2021 [21]	YOLOv3	SCDS	Monocular cameras, robot
2022 [11]	YOLO variants	UR Fall self-annotated RGB dataset	Video camera
2022 [25]	Modified YOLOv5s	URFD dataset	Microsoft Kinect cameras
2023 [14]	Multimodal data fusion, federated learning	UP-Fall dataset	Wearable sensors, cameras
2023 [40]	YOLOv5x, YOLOv5s.	CAUCA Fall	Webcam, IoT devices
2024 [22]	YOLOv7-fall, YOLOv7-tiny	Multiple cameras fall dataset, UR FD Dataset	RGB images

2.1. Machine learning assisted sensors based approaches

ML approaches are extensively utilized to predict falls in different environments [27]. For instance, an ML-based study [28] used visual frame data of URFD by utilizing various models and reported optimal performance using SVM and KNN. Another ML-assisted work proposed a 3-D axial accelerometer integrated with a wearable 6LowPAN device [29]. In order to achieve optimal performance in FD, the sensor data undergoes processing and analysis using a decision tree-based model. In the event of a detected fall, an alarm is triggered and the system quickly responds by sending messages to the authorized organizations that handle the situations. Furthermore, the work [30] used two datasets Le2i and the SEIN for FD along with the SVM. The events were discovered and the attributes were used as input for the classifier. Moreover, the work [31] proposed a robust ML-based FD algorithm for elders. The study proposed an ML method for detecting falls, with a focus on preserving privacy and reducing interference time [32]. The approach utilizes the encoding of skeletal features that represent the sequence of consecutive frames, along with the use of an SVM classifier. The researchers used a modified version of the cumulative sum (CUMSum) technique to combine the individual assessments made on the successive frames. They suggested employing the one-class classifier to recognize low-quality skeletons.

Moreover, the work [33] proposed an intelligent method using a sensor data approach. Their suggested network depends on accelerometer data collected from a wristwatch. However, achieving excellent precision is typically low since these devices are sensitive to interference resulting from various movements. Another investigation [34] used a gyroscope to measure acceleration and to recognize falls based on angular velocity. Both devices are employed to estimate the variation in values between leaping and falling rotations. In addition, the method described in [35] utilized the smartphone's integrated accelerometer, which was supposed to be located in the front pocket, to gather SAM, ZMean, and Xmean features. It is observable that they set the system so that if the values of SAM, Zeman, and Xmean exceed the threshold, a fall is automatically detected. The user's location is immediately transmitted to an alert center for prompt medical attention. In addition, the research [36] proposed an FD by keeping in view thresholding techniques. The researchers performed empirical analysis and investigations on four different kinds of falls and eight distinct kinds of activities of daily life using an integrated sensors setup that combines an inertial sensor and a plantar-pressure measuring unit. The FD efficiency was assessed by assessing the information collected using both the threshold approach and the decision-tree technique.

2.2. Deep learning assisted visual intelligence based approaches

In recent years, DL-driven visual intelligence-based techniques have become prevalent in FD [41]. Among these approaches, the YOLO versions have gained widespread adoption due to their superior real-time performance [42,43]. For instance, YOLOv2 was employed in [19] to detect humans in images, utilizing pre-trained weights. Afterward, the YOLOv2 model was modified by fine-tuning the final layer using their dataset, including 500 manually annotated images. Similarly, [20] proposed a YOLOv3-based FD method capable of detecting multiple individuals in frames. Another notable contribution [21] introduced an assisted system for fall risk management, utilizing monocular cameras and a humanoid robot. They suggested using a CV network to evaluate the level of disorder in rooms containing single-lens visual sensors in a place of residence. Furthermore, a socially accepted robo is involved to improve communication between the resident and the system. This

technology has been developed specifically for maximum adaptability, rendering it highly suitable for easy integration into existing intelligent settings. Furthermore, [11] proposed a YOLOv4-based network specifically tailored for real-world FD, utilizing the UR FD dataset containing approximately 1691 fall and 1731 normal samples. A fundamental feature of the suggested approach is that it requires no environmental sensors since it automatically recognizes falls in images from standard visual sensors. The system performed the extraction of features using the UR Fall self-annotated RGB dataset. Additionally, [25] presented an FD network based on enhanced YOLOv5s. The study proposes a real-time detection approach for detecting senior citizen fall behavior using a modified version of YOLOv5s. The goal was to quickly determine whether an old person has fallen, helping them to get prompt proper care. The initial phase involves replacing the conventional basic convolution with the asymmetric convolution blocks module in the Backbone network to enhance its capacity for obtaining optimal features. Next, they incorporated a spatial attention mechanism module into the residual structure of the Backbone network to improve the extraction of feature location information. The contributions are good in the structure of YOLOv5. However, the lack of a diverse dataset for training limited its performance and scalability. Different YOLO versions with progressive modifications are deployed in CV tasks [44]. Most recently, [22] introduced YOLOv7-fall with the claim of enhanced feature extraction capabilities and reduced model parameters. However, their training dataset also lacked high more diversity, consisting of only 4016 images. The limitations in the related literature collectively highlight the need for diverse data and progressive enhancements in the SOTA DL-driven visual intelligence-based approach for FD.

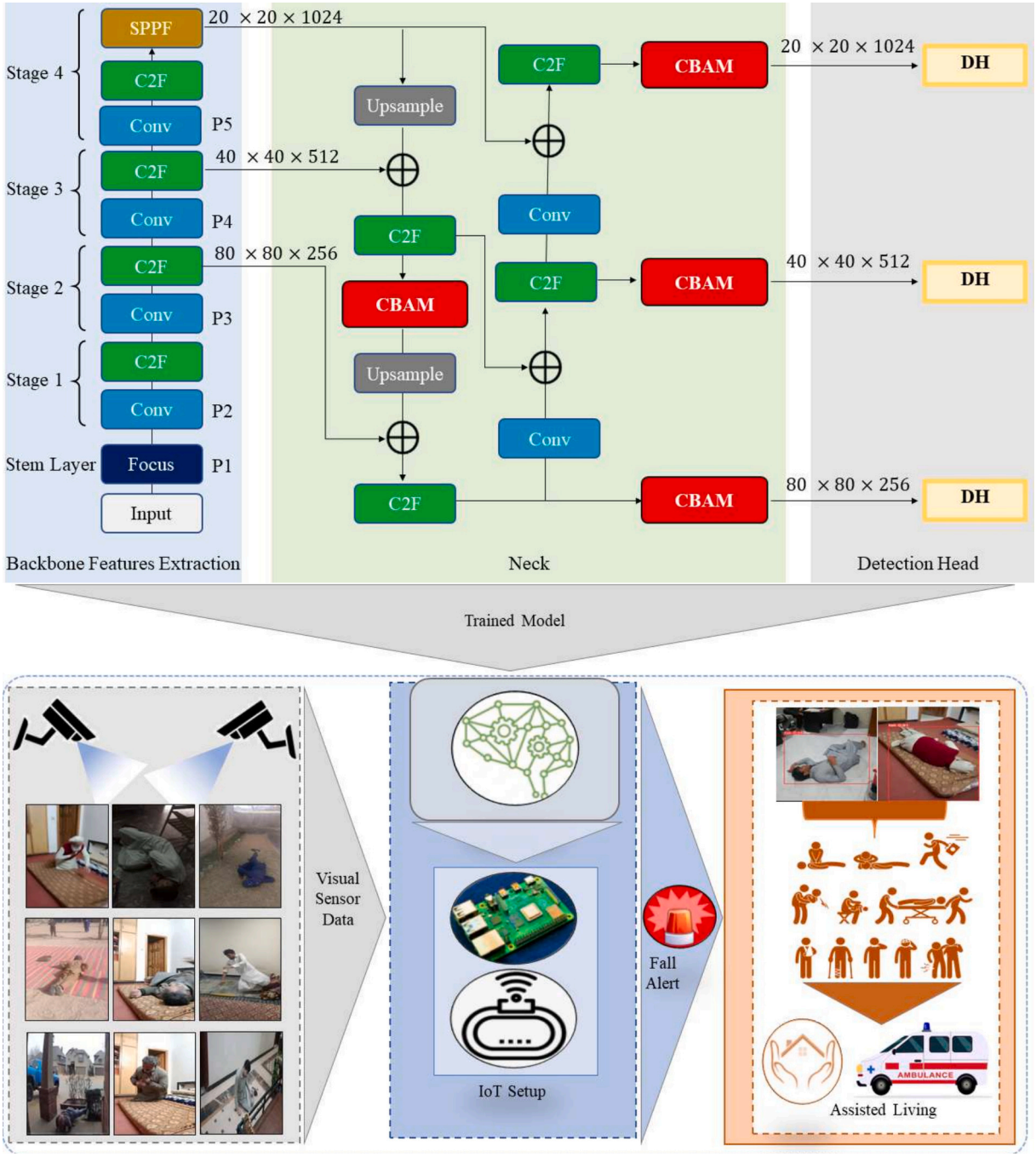


Fig. 1. Visual overview of the proposed FD network for assisted living of fallen persons.

3. Proposed methodology

This section provides an overview of the proposed benchmark dataset, the internal architecture of the proposed network, the feasibility study, and the modifications applied to YOLOV8S for improved FD. Fig. 1 illustrates the general overview of the training, testing, and applicability of the proposed network, which is further elaborated in the subsequent sub-section.

3.1. DiverseFALL10500: Diverse fall detection dataset

We significantly contributed to the FD domain by collecting and annotating a large-scale benchmark dataset. After an extensive study of the related literature, we included all the important aspects and factors needed to contribute to a challenging benchmark in the domain. We incorporated visual samples by considering key factors such as varying illumination, diverse environments, age diversity, different clothing and uniforms, and involving multiple participants in the data collection and annotation process. It is observable that the previous attempts at contributing datasets are valuable [45]. However, they often considered the inclusion of uniform frames as shown in Fig. 7 which limits the network to learn more in complex and uncertain visual environments. The diversity of our data is shown in visual samples in Fig. 2. Table 2 shows the inclusion of multiple aspects of diversity and key features in comparison with other datasets. Our dataset includes images from people of diverse ages and detailed annotation protocols with rigorous quality control measures. The data includes approximately 10,500 annotated images with 5806 samples of falls and 5132 instances of non-falls. We considered the following necessary aspects after considering the limitations of the literature to ensure the diverse collection of the database:

- **Various Backgrounds Inclusion:** Acknowledging fall incidents across diverse settings, our dataset integrates samples taken in indoor and outdoor environments, encompassing a spectrum of conditions and backgrounds. This deliberate inclusion of a representative array of environmental factors aims to confront the challenge associated with adapting FD models to diverse conditions.
- **Illumination Variations:** The dataset intentionally includes changes in lighting conditions to make the model detect falls at various levels of illumination. This feature can assist the model to function efficiently in a wide range of situations. The images are taken in different lighting conditions, such as artificial light, daylight, and low light.
- **Clothing and Attire Variability:** We ensured an extensive selection of clothing styles, colors, and diverse uniforms that the actors used to present a realistic and versatile visual looks. This particular illustration encourages an arrangement that is capable of recognizing fallen persons throughout various visual contexts, boosting its ability to change according to the demands of daily life.
- **Multi-Participant Involvement:** Multiple participants were involved in the dataset creation process, adding a layer of variability. This multiplicity reflects a collective understanding of different perspectives, contributing to a well-rounded dataset that captures the nuances of real-world fall incidents.

Fig. 2. Sample images of our newly proposed DiverseFALL10500 dataset for fall detection. The images show diverse representations of various environments.

Table 2
Comparison of datasets for human fall detection. The sign √ shows the incorporated features while the √ √ presents the high level of inclusion.

Year	Dataset	Sensors	Labels for YOLO	Variable Visual Appearances	Occlusion	Variety in fall angles	Light Conditions	Age Diversity	Clothing and Attire Variations	Background Diversity
2010	**Multiple cameras fall** [46]	**RGB**	–	–	√	–	**artificial**	–	–	√
2012	**Le2i** [47]	**RGB**	–	–	√	√	**natural, artificial**	–	–	–
2014	**SDUFall** [48]	**Kinect**	–	–	–	–	**natural, artificial**	–	–	–
2014	**EDF&OCCU** [10]	**Kinect**	–	–	√	–	**artificial**	–	–	–
2014	**UR FD** [49]	**Kinect**	–	–	–	–	**artificial**	–	–	√
2016	**FUKinect-Fall** [50]	**Kinect**	–	–	–	√	–	–	–	√
2017	**FD** [51]	**RGB Kinect**	–	–	–	√	**natural, artificial**	–	–	√
2019	**UPFall** [12]	**RGB**	–	–	–	√	**natural, artificial**	–	–	√
2022	**CAUCAFall** [45]	**RGB**	√	–	√	√	**natural, artificial, no light**	–	–	√
2024	**DiverseFALL10500 (Ours)**	**RGB**	√	√ √	√ √	√ √	**natural, artificial, night views**	√ √	√ √	√ √

3.2. Feasibility of fall detection network

Object detection methods have become pivotal in various image-related tasks [52,53], with recent advancements enabling efficient localization of regions of interest. Particularly in FD, where precise object localization and detection are crucial, existing methodologies face challenges in balancing computational demands with performance. The YOLO variants have emerged as prominent solutions, offering real-time detection capabilities with enhanced performance [54]. Notably, the evolution of YOLO models, from YOLOv1 to YOLOv8, witnessed progressive modifications aimed at improving detection accuracy and reducing model complexity [55]. In contrast to earlier YOLO versions till version 4, YOLOv5 demonstrates improved detection performance, low computational complexity, and model size, making it a practical choice for resource-constrained devices. Each of these models refers to a unified structure of networks that incorporates multiple key elements necessary for optimal detection of objects. The layer that receives input includes the role of preprocessing images by resizing them to a specific size and leveling the pixel values. This takes place to make it simpler to perform further processing. Additionally, the backbone unit provides the fundamental foundation for obtaining features from input images, using feature pyramid networks (FPN) for capturing representations at various sizes. Furthermore, the neck module includes a sequence of layers specially designed to combine data from different levels of the backbone network, allowing the detailed feature representation. The prediction using head components plays an important role in generating accurate predictions, involving bounding box coordinates, confidence scores, and class probabilities. These architectural components interact seamlessly to provide strong detection of objects, as described in research investigations [56]. The YOLOv8 was announced as a recent enhancement to the YOLO family. SOTA object detection models are highly promising and feasible for detection tasks [57] due to their optimal performance and real-time processing capabilities. However, they require adaptation and fine-tuning on specialized diverse datasets to accurately grasp falls even in challenging environments. Our investigation aims to boost the accuracy and robustness of FD systems by improving the feasible modifications.

3.3. YOLOV8 network architecture

YOLOv8 demonstrates exceptional detection performance and it adopts a unique design framework that avoids the use of anchor boxes [58]. It presents modifications in its elements compared to its prior versions, with a majority of modules adopting the framework of YOLOv5. The structure of YOLOv8 can be categorized into three separate components: Backbone, Neck, and detection by head part, as seen in Fig. 1. It is observable that YOLOv8 holds several important features from YOLOv5, including the Contextual Spatial Pyramid (CSP) and the Path Aggregation Network with FPN (PAN-FPN). However, YOLOv8 additionally makes major enhancements that set it distinct. In the Backbone component of YOLOv5, the C3 module is replaced with a single C2F module, but the rest of the modules generally remain consistent with YOLOv5. The Neck component incorporates the PANet configuration, which streamlines the vertical link by using two C2Fs. Additionally, it maintains two C2Fs and two CBSs for the left and right connections, resulting in a simplified structure for the Neck region. In addition, the C3 module gets replaced with a C2f module, which boosts the performance of the original 6×6 convolution in the backbone by refining it to a more effective 3×3 convolution. Furthermore, the head part utilizes an approach by segregating the detection and classification heads [59]. We have optimized the YOLOV8S framework for precise as well as efficient FD, overcoming the limitations of earlier research, and taking motivation from enhancements regarding efficiency and inference time, especially in real-world applications. The solution that we provide offers higher precision, a better capacity to detect in real-time, flexibility, fewer false alarms, and lower computing complexity. Additionally, it provides the ability to detect falls in challenging environments as included in the proposed dataset.

3.4. Optimized proposed network

We introduce two progressive modifications designed to achieve a high level of performance. More specifically, we incorporated the focus module at the starting of the backbone to assist intermediate features. Moreover, CBAMs areintegrated into YOLOV8S to boost and refine the features before ultimate predictions. The aforementioned improvements are further detailed in the following subsections.

3.4.1. Integration of focus module

In our network, we replaced the Conv block with the Focus module in the backbone to enhance the feature extraction process. We incorporated the Focus module into the backbone to strengthen its extraction of features for more accurate FD. It is observable that it initially appeared in the YOLOv5 architecture [60]. The Focus module enhances the model's capacity for integrating spatial aspects by splitting the input into four segments and merging them along the channel axis as shown in

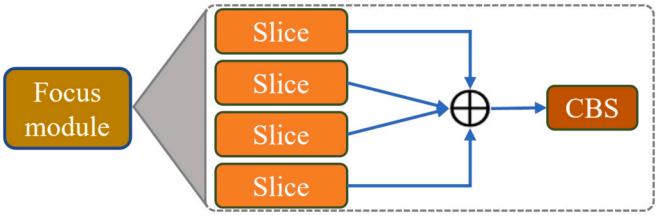


Fig. 3. Overview of the Focus module in the backbone.

Fig. 3. In contrast to the Conv block, which may neglect complex spatial details, the Focus module ensures that the following layers gain a more comprehensive spectrum of information. This attempt enables improved learning and generalization, particularly for intricate tasks such as FD. It divides the high-quality feature maps, distinguishing them from traditional image downsampling techniques. The downsampled image is subsequently transmitted into the backbone network. By slicing the image, every pixel is given a value, similar to a nearest-neighbor downsampling technique. This preserves all of the initial information while minimizing the dimensional parameters of the channel space. After the slicing process, the downsampled image goes through a convolutional execution, generating a feature map that is double the dimension of the downsampled result yet preserves all of the original information. Following that, the feature map is subsequently transmitted to the next convolutional layer using batch normalization and an activation function. The conducted empirical analysis on the YOLOV8S network indicates that the Focus unit improves the accuracy of the network. In addition, the Focus module effectively manages computational efficiency for faster inference with no mAP penalty.

3.4.2. *Attention over attention*

Attention mechanisms have demonstrated significant improvements in performance across a range of visual intelligence tasks by effectively refining relevant features [61,62]. We proposed integration of CBAMs in the YOLOV8S architecture enhances its ability to focus on relevant features by incorporating both spatial and channel attention mechanisms. The backbone of the network begins with an early focus module and then conv layers that gradually decrease the spatial dimensions while simultaneously increasing the depth of the feature maps. After these layers, other C2f modules carry more processing on the feature maps at different resolutions. Additionally, there is a spatial pyramid pooling layer capturing features. CBAMs are incorporated at multiple points in the architecture to enhance the quality of the feature maps. Initially, there are four stages in the backbone feature extraction module. We utilized features from stages 2, 3, and 4, which have dimensions with increasing numbers of channels and decreasing spatial resolutions. In the neck module, we applied CBAMs along with other convolution operations to these features as illustrated in Fig. 1. Finally, we obtained three output features and passed them directly to the detection heads. The final detection layer utilizes these enhanced features and integrates the results from three separate scales to provide precise predictions. The use of CBAMs ensures that the network can effectively emphasize key features while suppressing irrelevant ones, resulting in improving detection performance without a substantial increase in computational complexity. Fig. 4 demonstrates the integration of channel and spatial attention processes with skip connections.

3.4.2.1. Channel attention. In the context of object detection and classification tasks, CA helps the model to focus on channels that carry more relevant information while suppressing less informative ones. CA typically operates through a series of computational steps. First, it computes global average pooling and max pooling operations across each channel of the feature map. These pooling operations aggregate channel-wise, capturing the overall importance of each channel in representing discriminative features. The resulting pooled values are then passed through a shared multi-layer perceptron (MLP), which consists of fully connected layers with nonlinear activations. The MLP enhances the representation power by learning complex relationships between channels. A sigmoid activation function processes the output of the MLP. This transformed output serves as an attention map indicating each channel's importance. The attention map is applied element-wise to the original feature map during inference. Channels are scaled according to their corresponding attention values, effectively boosting informative channels while attenuating less useful ones. By incorporating CA into the network architecture, the model gains the capability to adaptively allocate computational resources to channels that contribute most significantly.

3.4.2.2. Spatial attention. SA is a critical component in enhancing the spatial awareness of neural networks by selectively highlighting informative regions within feature maps. In tasks like, SA enables the model to focus on spatial locations that are most relevant for accurately identifying and localizing objects of interest. The SA mechanism typically operates through a sequence of computational steps. Initially, it computes global max pooling and average pooling operations across the channel dimension of the input feature maps. These pooling operations aggregate spatial contexts, capturing the importance of each spatial location in representing discriminative features. Following pooling, SA applies a convolutional operation to the concatenated output of the pooled features. This convolutional layer acts as a spatial filter, learning to weight different spatial locations based on their importance. The output of this convolutional layer is an attention map that assigns higher values to significant spatial regions and lower values to less relevant ones. During inference, the attention map generated by SA is multiplied element-wise with the original feature map. This scaling operation effectively enhances the representation of informative spatial locations while suppressing noise and irrelevant background regions.

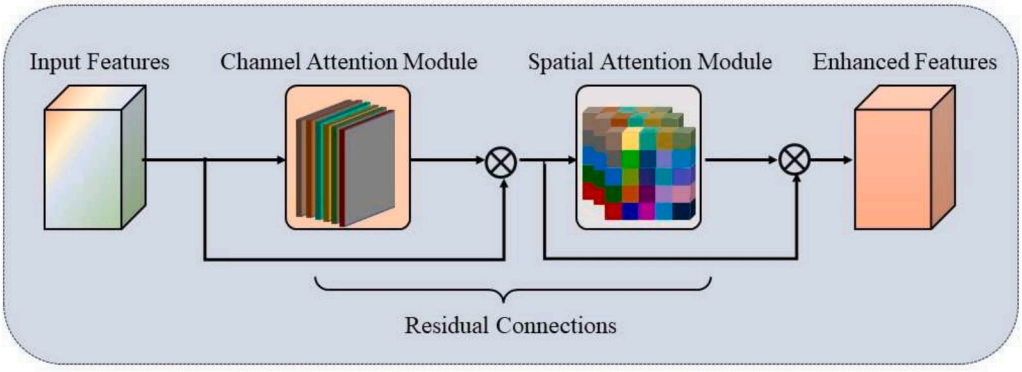


Fig. 4. Visual illustration of the CBAM in YOLOV8S.

Furthermore, we fuse the ultimately refined features with the intermediate features to provide progressive assistance. By integrating SA into the network architecture, as depicted in our proposed network in Fig. 4, the model gains the capability to adaptively focus on critical spatial details crucial for precise FD.

4. Results and discussion

In this section, we offer an extensive breakdown of the experimental setup, including hyperparameters, evaluation metrics, modules integration, ablation study, quantitative and qualitative results with computational complexity analysis, and ultimately, the comparison of our network's performance with different versions. Different evaluation metrics, including accuracy, precision, recall, F1-score, and mAP are used to assess the performances of the conducted experiments. The Precision-recall curve is visualized in Fig. 5 while the F1-confidence curve can be seen in Fig. 6. It is important to note that the visualized results are derived from a single split of the data. The final results represent the average performance across five different splits, as specified in the ablation study.

4.1. Experimental setting and hyper-parameter selection

The experiments were conducted on a system equipped with an Intel (R) Core(TM) i9-10,900X CPU @ 3.70GHz, featuring 10 cores and 192 gigabytes of random access memory. Training of the models was facilitated using an NVIDIA GeForce RTX 4090 graphical processing unit with 24 gigabytes of onboard memory. The development and training of proposed network was executed within the PyTorch framework on a Windows 10 platform. Several essential libraries were used to facilitate network development and training, including Seaborn, tqdm, NumPy, Scikit-learn, Pandas, Matplotlib, and Pillow. Various configurations of epochs, batch size, optimizer, and learning rate were explored during the conducted experiments. The most optimal results were obtained using 100 epochs and an input image size of 640 × 640. A batch size of 10 was selected to balance computational efficiency and network performance. For optimizer selection, Stochastic Gradient Descent (SGD) was chosen as experimentally validated, and specific hyperparameters were fine-tuned to enhance network convergence and performance. Experimental results and related literature in the FD domain came together to establish these hyperparameters. It is worth mentioning that the SGD optimizer was set up with the specific parameters: Momentum = 0.937, learning rate = 0.001 (after empirical validation), and Weight decay = 0.0005. The thorough choice of the conditions for experimentation and hyperparameters played an important role in attaining the optimal possible accuracy of the suggested network.

4.2. Datasets utilized and splitting

This section focuses on the datasets that were utilized in our research and the approach we applied to divide them into three categories: training, validation, and testing sets. The proposed DiverseFALL10500 dataset includes a diverse collection of images showing falls and normal complex images taken in different real-life environments. The proposed large-scale benchmark dataset presents a comprehensive representation of various environmental conditions as elaborated in the Table 2. The inclusion of various environmental and illumination conditions is efficient in ensuring the training of networks on diverse data. The data possesses a total of 10,500 images, with each image having an appropriate annotation as discussed in section 3.1. During the training phase, 70% of the data is fed to train network, 20% goes for validation, and the remaining 10% is set aside for testing. The second database CAUCAFall comprises 10 people contributing to falls and activities of daily living inside a real-life domestic location. The dataset features the inclusion of various environmental variables like changes in illumination and various textures. The data is organized into folders that are detailed to different subjects and each group has systematic labeling. The dataset is a good contribution to the FD domain to train new systems for detecting falls. However, the homogenous frames could limit the generalization of the model. Similar to DiverseFALL10500, this data is divided according to the established criteria for training, validation, and testing.

4.3. Evaluation parameters

The assessment of the proposed network is based on essential object detection metrics such as mAP, precision (Pr), recall (Re), and F1-score (F1). These metrics are well acknowledged in the field of object detection and are precisely specified mathematically in the following equations, as explained in [22].

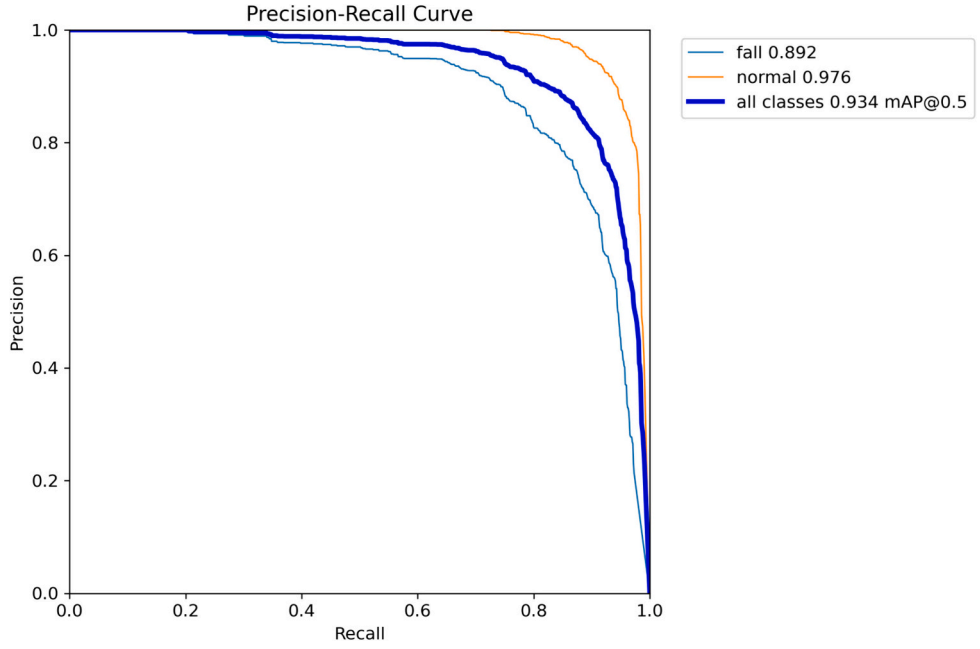


Fig. 5. Precision-recall curve of the proposed network.

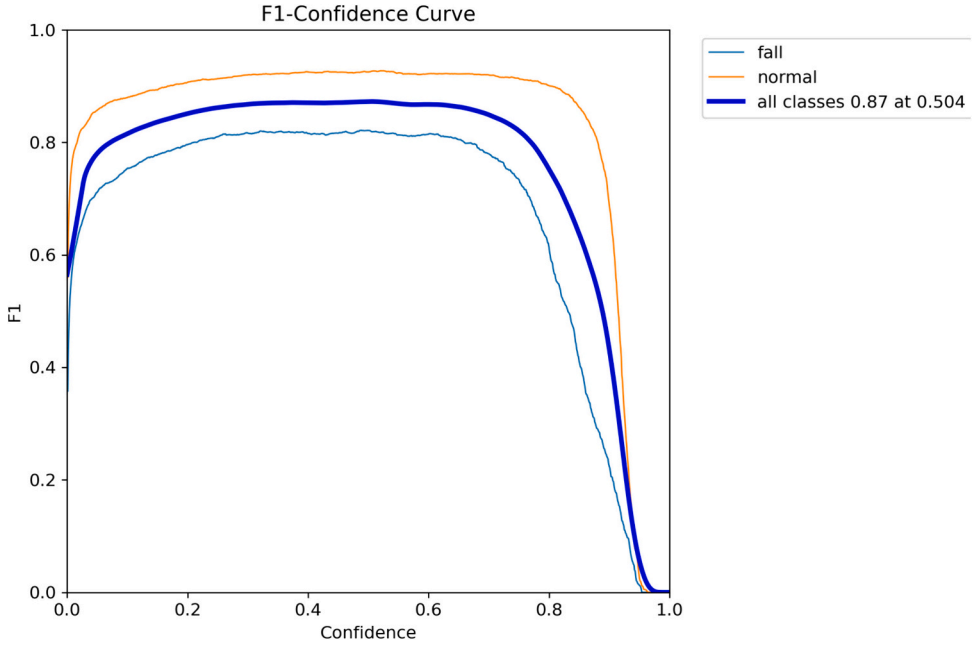


Fig. 6. F1-Confidence curve of the proposed network.

$$Pr = \frac{TP}{TP + FP} \tag{1}$$

$$Re = \frac{TP}{TP + FN} \tag{2}$$

TP denotes true positives, meaning correct predictions; FP defines false positives, which signifies an incorrect identification; and FN indicates false negatives, showing a failure to identify.

$$F1 = 2 \times \left(\frac{\mathrm{Pe} \times Re}{\mathrm{Pe} + Re}\right) \tag{3}$$

$$AP = \sum_{i=1}^{c} Pr(j) \times \Delta Re(j) \tag{4}$$

$$mAP = \frac{1}{2} \sum_{i=1}^{n} AP_i \tag{5}$$

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 related to the model's performance, effectiveness, and practical feasibility in real-world scenarios.

4.4.1. Ablation studies

We practiced an ablation study to assess the result of different components in our proposed network. Table 4 presents the findings of our ablation investigation, exhibiting the efficiency measurements with modifications that incorporate certain modules. Upon examining the proposed dataset, we see a consistent improvement in performance indicators as we include additional modules into the fundamental YOLOV8S structure. Initially, the base YOLOV8S network achieves an mAP of 0.920. Upon incorporating the focus module into the backbone, denoted as YOLOV8S + Focus, we observe a slight improvement in mAP to 0.923. Subsequently, integrating the CA module into the neck enhances performance, resulting in an mAP of 0.927. With the integration of the SA module, the network achieves an mAP of 0.928. Finally, our complete model, which incorporates both the focus and the CBAM module at the specified stages, achieves the highest mAP of 0.935, demonstrating the synergistic effect of these modules on FD accuracy. Similar trends are observed when evaluating the CAUCAFall dataset, with incremental improvements in performance as additional modules are integrated. The base YOLOV8S model achieves an mAP of 0.9950, which increases marginally to 0.9951 by adding the focus module. Integrating the CA module further yields the same mAP of 0.9951. Finally, our complete proposed network achieves the highest mAP of 0.9954, underscoring the effectiveness of incorporating both the focus and CBAM modules. It is important to mention that to ensure reliable and robust results, we performed experiments using five distinct random splits of the proposed dataset. Each division was used for both training and testing our network. We presented the average performance of these studies to provide a thorough assessment of the accuracy of the network. These results underscore the importance of each module in contributing to the overall performance of our network. The ablation study provides valuable insights into the individual and combined effects of these modules on FD accuracy, validating the efficacy of our proposed enhancements.

4.4.2. Quantitative evaluations

In the quantitative analysis, we provide two types of results which include the comprehensive comparison of various SOTA FD models, shedding light on their performance across two distinct datasets: Our developed Diversefall10500 and CAUCAFall dataset [45]. Our analysis incorporates a comprehensive assessment of well-established baseline models, where we thoroughly reimplemented 13 different detection models to evaluate the effectiveness of our methodology. Table 3 presents a detailed breakdown of key metrics, including mAP, Precision, F1-score, and Recall, for each model on both datasets. Among the models evaluated, our proposed modification to the YOLOV8S architecture stands out, achieving the highest mAP of 0.935 on the proposed dataset and an impressive 0.9954 on the CAUCAFall benchmark. These results significantly surpass those of other models, underscoring the efficacy of our approach in enhancing FD performance. It is observable that the superior performance of the CAUCAFall dataset is evident from the tabulated results. However, it can be visually seen from Fig. 7 that the recorded visual samples often exhibit a high degree of homogeneity.

Table 3
Quantitative analysis of different SOTA object detection models using the proposed FD and benchmark CAUCAFall datasets.

S No	Model	DiverseFALL10500				CAUCAFall			
		mAP	Precision	F1-score	Recall	mAP	Precision	F1-score	Recall
1	Faster R-CNN	0.831	0.837	0.813	0.809	0.9903	0.9891	0.9902	0.9924
2	yolov3	0.842	0.848	0.827	0.809	0.9912	0.9904	0.9916	0.9931
3	yolov4	0.825	0.818	0.801	0.794	0.9896	0.9901	0.9898	0.9913
4	yolov5n	0.870	0.810	0.839	0.852	0.9924	0.9921	0.9935	0.9942
5	yolov5s	0.906	0.895	0.878	0.845	0.9932	0.9943	0.9954	0.9961
6	yolov5m	0.848	0.850	0.825	0.807	0.9911	0.9914	0.9923	0.9927
7	yolov5l	0.858	0.805	0.813	0.837	0.9926	0.9925	0.9934	0.9942
8	yolov5x	0.834	0.769	0.797	0.829	0.9898	0.9892	0.9903	0.9906
9	yolov8n	0.863	0.810	0.822	0.837	0.9925	0.9921	0.9923	0.9935
10	YOLOV8S	0.920	0.895	0.879	0.851	0.9950	0.9977	0.9976	0.9978
11	yolov8m	0.899	0.880	0.868	0.851	0.9949	0.9948	0.9952	0.9950
12	yolov8l	0.871	0.855	0.856	0.845	0.9947	0.9941	0.9914	0.9932
13	yolov8x	0.903	0.885	0.874	0.853	0.9945	0.9937	0.9941	0.9953
14	**Proposed network**	**0.935**	**0.900**	**0.884**	**0.859**	**0.9954**	**0.9982**	**0.9981**	**0.9984**

Fig. 7. Sample images of CAUCAFall dataset.

This homogeneity poses a limitation on the model's ability to generalize effectively, particularly in diverse environmental and background conditions. It's noteworthy that while several models, such as YOLOv5s and YOLOv8m, demonstrate competitive performance, they fall short of our proposed network in terms of mAP and other evaluation metrics. This highlights the importance of the modifications we introduced, including the integration of the focus module in the backbone and the progressive integrations of CBAMs, which contribute to the superior performance. Our choice of YOLOV8S as the base model for incorporating these modifications was strategic, considering its strong performance as well as its suitability for real-time FD applications. The robustness and efficiency of the YOLOV8S architecture provide a solid foundation for integrating advanced techniques aimed at improving FD accuracy. The impacts of these findings are substantial for the advancement of FD systems, especially within medical settings where immediate action could minimize the possibility of major damages.

Table 4
Ablation study of our network with the integration of different modules.

	Model	mAP	Precision	F1-score	Recall
Proposed Dataset	YOLOV8S	0.920	0.895	0.879	0.851
	YOLOV8S + Focus	0.923	0.896	0.879	0.852
	YOLOV8S + CA	0.927	0.898	0.880	0.855
	YOLOV8S + SA	0.928	0.899	0.883	0.858
	Proposed network	**0.935**	**0.90**	**0.884**	**0.859**
CAUCAFall	YOLOV8S	0.9950	0.9977	0.9976	0.9978
	YOLOV8S + Focus	0.9951	0.9978	0.9976	0.9979
	YOLOV8S + CA	0.9951	0.9979	0.9978	0.9980
	YOLOV8S + SA	0.9953	0.9978	0.9980	0.9981
	Proposed network	**0.9954**	**0.9982**	**0.9981**	**0.9984**

4.4.3. Qualitative analysis

We present the visualized outcomes from two distinct angles: First, we exhibit the performance of our network on both datasets, illustrated in Fig. 8. Secondly, we offer a visual comparative analysis with existing models, as shown in Fig. 9. It is noteworthy that the first three visual samples in each row of Fig. 9 depict falls, while the remaining two depict normal instances. Our network effectively grasps fall regions in images, exhibiting a commendable mAP score. For evaluating network performance, we compare our results with four SOTA models: Faster R-CNN, YOLOv3, YOLOv5, and YOLOv8. Notably, our network achieves superior performance with detection scores reaching 82%, 79%, 83%, 85%, and 88% mAP values on Faster R-CNN, YOLOv3, YOLOv5, YOLOv8, and our network, respectively, as illustrated in the first column of Fig. 9. The results indicate that while Faster R-CNN outperforms YOLOv3, it comes with higher computational costs. In the second column, our network surpasses Faster R-CNN, YOLOv3, YOLOv5, and YOLOv8, exhibiting higher detection scores, surpassing them by 16%, 20%, 13%, and 11%,

Table 5
Comparison of five different YOLOv8 variants using various optimizers and learning rates experimented on our proposed dataset.

Learning Rate (0.001)						Learning Rate (0.0001)			
OP	Model	mAP	Precision	F1-score	Recall	mAP	Precision	F1-score	Recall
	YOLOv8n	0.907	0.862	0.869	0.841	0.894	0.868	0.860	0.840
	YOLOV8S	0.881	0.879	0.853	0.837	0.901	0.890	0.859	0.840
Adam	YOLOv8m	0.892	0.851	0.834	0.819	0.913	0.847	0.874	0.847
	YOLOv8l	0.905	0.838	0.868	0.850	0.898	0.892	0.874	0.846
	YOLOv8x	0.919	0.865	0.861	0.820	0.912	0.845	0.864	0.830
	YOLOv8n	0.864	0.853	0.858	0.816	0.884	0.845	0.857	0.847
	YOLOV8S	0.894	0.876	0.871	0.843	0.906	0.884	0.876	0.834
AdamW	YOLOv8m	0.916	0.887	0.864	0.838	0.864	0.857	0.852	0.839
	YOLOv8l	0.903	0.848	0.843	0.849	0.916	0.864	0.867	0.849
	YOLOv8x	0.913	0.894	0.881	0.856	0.916	0.885	0.877	0.847
	YOLOv8n	0.884	0.883	0.876	0.837	0.903	0.885	0.878	0.832
	YOLOV8S	0.904	0.872	0.874	0.841	0.916	0.897	0.863	0.848
Nadam	YOLOv8m	0.915	0.880	0.858	0.849	0.902	0.893	0.880	0.835
	YOLOv8l	0.877	0.873	0.867	0.842	0.902	0.853	0.872	0.844
	YOLOv8x	0.916	0.865	0.867	0.843	0.901	0.872	0.857	0.825
	YOLOv8n	0.881	0.832	0.837	0.847	0.853	0.893	0.865	0.844
	YOLOV8S	0.893	0.875	0.873	0.832	0.905	0.876	0.878	0.849
Radam	YOLOv8m	0.912	0.883	0.877	0.851	0.892	0.864	0.849	0.851
	YOLOv8l	0.902	0.871	0.876	0.852	0.914	0.882	0.850	0.834
	YOLOv8x	0.914	0.872	0.875	0.842	0.893	0.875	0.871	0.831
	YOLOv8n	0.882	0.872	0.856	0.842	0.881	0.893	0.860	0.833
	YOLOV8S	0.873	0.882	0.851	0.823	0.881	0.892	0.868	0.824
RMSProp	YOLOv8m	0.883	0.862	0.869	0.831	0.844	0.871	0.863	0.844
	YOLOv8l	0.882	0.863	0.866	0.828	0.883	0.861	0.873	0.846
	YOLOv8x	0.901	0.894	0.875	0.844	0.892	0.883	0.861	0.844
	YOLOv8n	0.882	0.855	0.859	0.847	0.891	0.887	0.849	0.839
	YOLOV8S	0.920	0.895	0.879	0.851	0.919	0.897	0.879	0.852
SGD	YOLOv8m	0.875	0.893	0.857	0.821	0.914	0.895	0.871	0.849
	YOLOv8l	0.883	0.862	0.873	0.824	0.892	0.872	0.868	0.842
	YOLOv8x	0.882	0.883	0.863	0.842	0.883	0.884	0.867	0.853
Proposed network with SGD		**0.935**	**0.90**	**0.884**	**0.859**	**0.928**	**0.899**	**0.878**	**0.855**

Table 6
Comparing five different YOLOv8 variants using various optimizers and learning rates experimented on CAUCAFall benchmark.

Learning Rate (0.001)						Learning Rate (0.0001)			
OP	Model	mAP	Precision	F1-score	Recall	mAP	Precision	F1-score	Recall
	YOLOv8n	0.9948	0.9958	0.9966	0.9975	0.9948	0.9981	0.9966	0.9952
	YOLOV8S	0.9947	0.9954	0.9966	0.9968	0.9946	0.9956	0.9967	0.9978
Adam	YOLOv8m	0.9948	0.9972	0.9969	0.9965	0.9945	0.9963	0.9967	0.9971
	YOLOv8l	0.9948	0.9971	0.9966	0.9961	0.9945	0.9970	0.9968	0.9966
	YOLOv8x	0.9944	0.9950	0.9959	0.9969	0.9937	0.9943	0.9957	0.9971
	YOLOv8n	0.9947	0.9971	0.9966	0.9960	0.9946	0.9973	0.9967	0.9961
	YOLOV8S	0.9946	0.9957	0.9965	0.9974	0.9947	0.9967	0.9965	0.9963
AdamW	YOLOv8m	0.9944	0.9958	0.9965	0.9972	0.9941	0.9960	0.9965	0.9969
	YOLOv8l	0.9943	0.9972	0.9974	0.9975	0.9942	0.9971	0.9972	0.9973
	YOLOv8x	0.9937	0.9950	0.9960	0.9970	0.9931	0.9944	0.9961	0.9979
	YOLOv8n	0.9947	0.9966	0.9972	0.9977	0.9945	0.9967	0.9970	0.9973
	YOLOV8S	0.9945	0.9956	0.9961	0.9966	0.9948	0.9959	0.9963	0.9967
Nadam	YOLOv8m	0.9946	0.9961	0.9961	0.9960	0.9945	0.9970	0.9964	0.9957
	YOLOv8l	0.9944	0.9946	0.9961	0.9977	0.9943	0.9964	0.9962	0.9959
	YOLOv8x	0.9944	0.9965	0.9964	0.9963	0.9939	0.9948	0.9958	0.9969
	YOLOv8n	0.9947	0.9962	0.9969	0.9976	0.9946	0.9980	0.9974	0.9967
	YOLOV8S	0.9945	0.9956	0.9961	0.9966	0.9946	0.9949	0.9962	0.9975
Radam	YOLOv8m	0.9945	0.9962	0.9959	0.9956	0.9944	0.9944	0.9962	0.9981
	YOLOv8l	0.9943	0.9964	0.9965	0.9966	0.9941	0.9956	0.9966	0.9976
	YOLOv8x	0.9943	0.9966	0.9966	0.9966	0.9939	0.9940	0.9956	0.9973
	YOLOv8n	0.9942	0.9919	0.9936	0.9953	0.9946	0.9944	0.9946	0.9948
	YOLOV8S	0.9949	0.9956	0.9942	0.9929	0.9944	0.9935	0.9943	0.9951
RMSProp	YOLOv8m	0.9945	0.9955	0.9944	0.9933	0.9948	0.9963	0.9952	0.9942
	YOLOv8l	0.9949	0.9963	0.9957	0.9951	0.9949	0.9962	0.9968	0.9973
	YOLOv8x	0.9946	0.9896	0.9927	0.9958	0.9949	0.9957	0.9969	0.9981
	YOLOv8n	0.9947	0.9962	0.9956	0.9949	0.9947	0.9965	0.9956	0.9948
	YOLOV8S	0.9950	0.9977	0.9976	0.9978	0.9952	0.9981	0.9974	0.9981
SGD	YOLOv8m	0.9942	0.9938	0.9950	0.9962	0.9942	0.9924	0.9952	0.9980
	YOLOv8l	0.9949	0.9957	0.9964	0.9971	0.9949	0.9954	0.9967	0.9981
	YOLOv8x	0.9949	0.9970	0.9974	0.9976	0.9949	0.9969	0.9974	0.9979
Proposed network with SGD		**0.9954**	**0.9982**	**0.9981**	**0.9984**	**0.9953**	**0.9982**	**0.9977**	**0.9981**

Table 7
A comparison of the proposed network with different YOLOv8 versions, considering GFLOPs, parameters, FPS, and model size (MS), mAP of the proposed dataset represents (mAP (1) and (2).

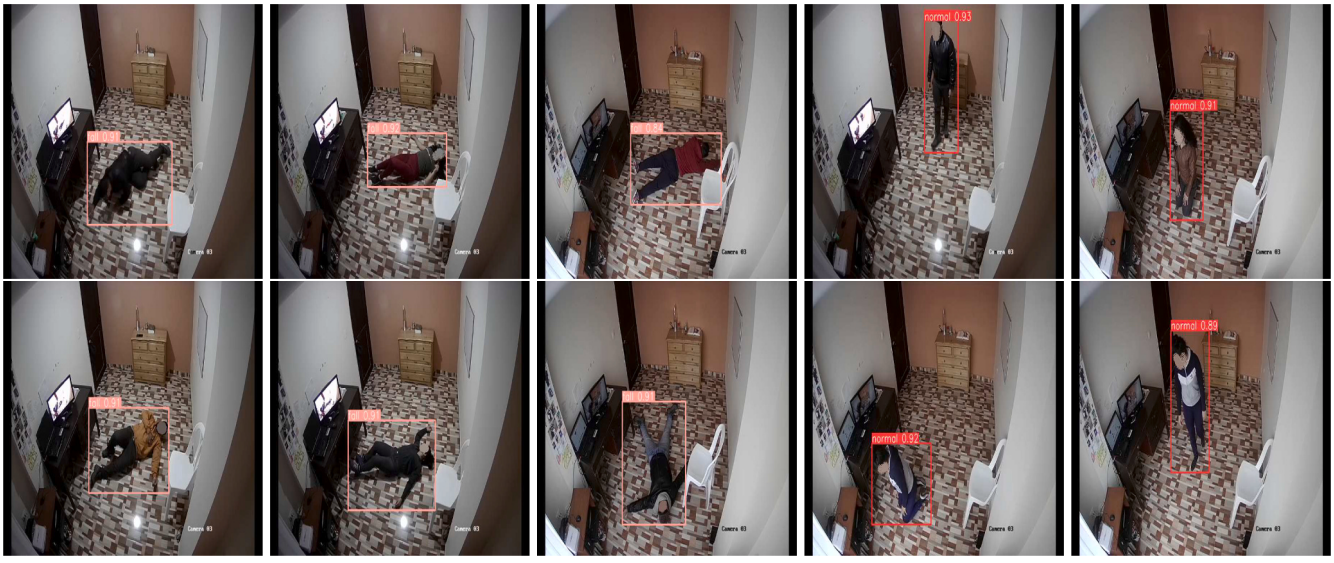
Methods	GFLOPs	Parameters	FPS	MS	mAP (1)	mAP (2)
YOLOv8x	258.5	68,229,648	83.81	130.53	0.903	0.676
YOLOv8I	165.7	43,691,520	69.74	83.70	0.871	0.652
YOLOv8m	79.3	25,902,640	98.27	49.70	0.899	0.633
YOLOV8S	28.82	11,166,560	130.92	21.53	0.920	0.718
YOLOv8n	8.86	3,157,200	117.99	6.23	0.863	0.584
Proposed network	**7.30**	**11,550,110**	**126.00**	**22.22**	**0.935**	**0.744**

respectively. Similarly, in the third column of Fig. 9, our network achieves a higher detection score compared to the runner-up YOLOv8. Moreover, in the fifth column of Fig. 9, our network outperforms the runner-up YOLO models with a marginal increase of 4% on normal instances. These visual illustrations highlight that the proposed network grasps the falls and normal patterns effectively even in challenging environments.

4.4.4. Technical analysis of optimizers, and learning rates

The tabulated results (Tables 5 and 6) offer a comprehensive comparison of various YOLOv8 variants trained on both datasets, along with our proposed network. Contrary to the common assumption that larger model variants consistently outperform smaller ones, our findings challenge this notion. Despite the larger size of YOLOv8x, our network consistently outperforms it across all evaluated metrics for both

A



B

Fig. 8. Qualitative samples taken from some of the challenging results of our network outputs using the proposed dataset (A) and CUCAFall dataset (B), with bounding boxes annotated by class name and confidence scores.

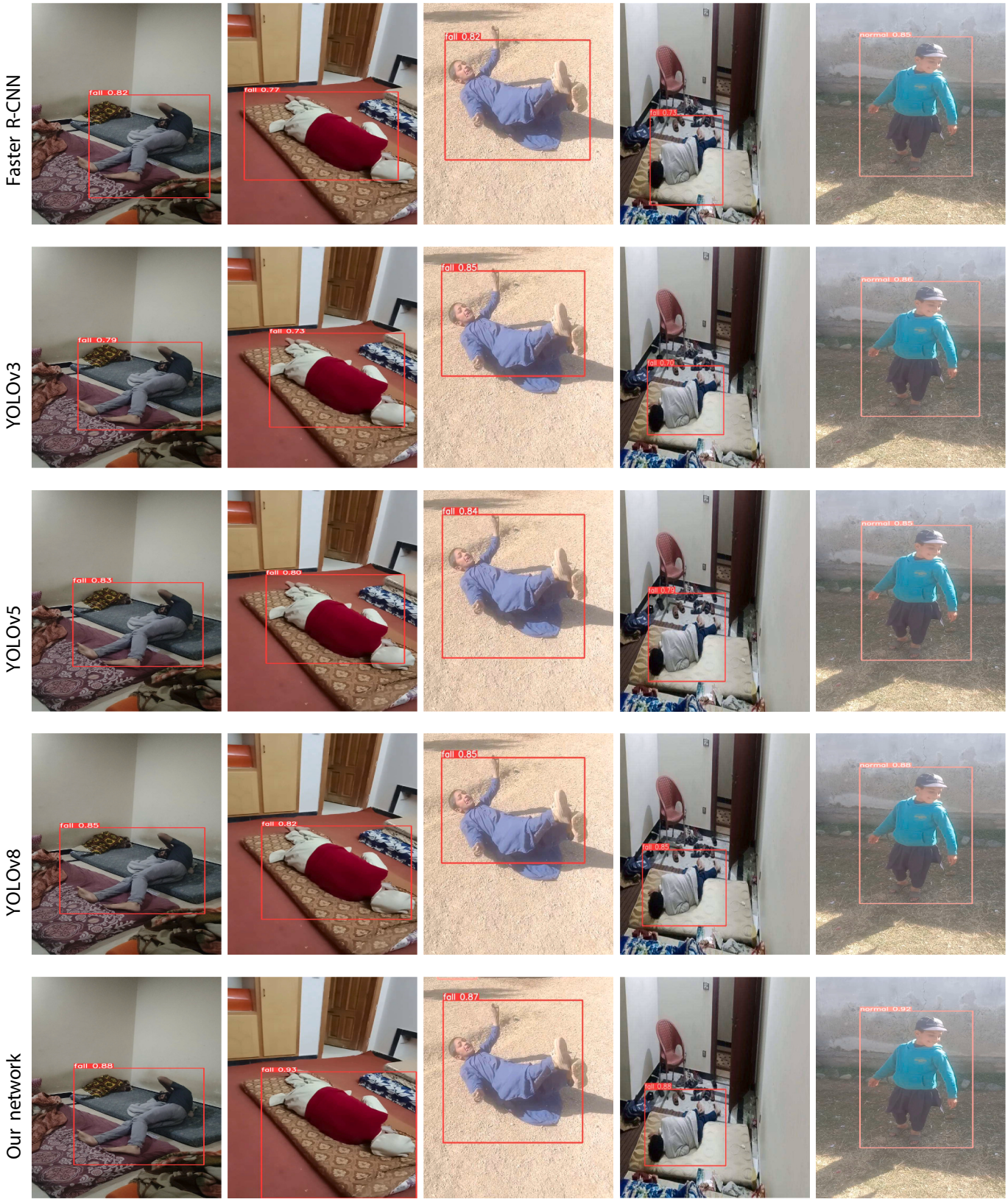


Fig. 9. A visualized comparison among the proposed network, YOLOv8, YOLOv5, YOLOv3, and Faster R-CNN over our proposed dataset to assess the robustness of the network.

datasets, including mAP, Precision, F1-score, and Recall. This suggests that factors beyond model size, such as architecture design and training strategy, significantly influence performance in object detection tasks across diverse datasets. The choice of optimizer emerges as a critical factor influencing model performance, especially when considering the nuances of different datasets. While certain optimizers may exhibit superior performance for specific YOLOv8 variants and learning rates on the proposed dataset, the trend may differ for the other dataset. However, our network demonstrates robustness to optimizer choice, consistently achieving remarkable results with the suggested optimizer used. The findings clearly demonstrate the influence of the learning rate on network performance. Generally, greater learning rates result in better results on both datasets, as seen by the consistent upward trend noticed across several metrics. Our proposed network surpasses the different versions of YOLOv8 on both datasets, highlighting the significance of our network across various data distributions. These results enhance understanding of performance and provide valuable insights for selecting suitable structures, optimization methods, and learning rates that are customized to specific datasets.

4.4.5. Computational complexity

The computational complexity of the proposed network with performance metrics, is presented in Table 7. This tabulated results compares our network with different versions of YOLOv8 in terms of GFLOPs, parameters, FPS, model size (MS), and mAP. Our network highlighted in bold demonstrates competitive computational efficiency with a GFLOPs value of 7.30, indicating efficient utilization of the integrated modules to strike a balance between computational resources and optimal performance. Despite this efficiency, our network maintains a relatively high level of mAP on both datasets, outperforming all other YOLOv8 variants. It is observable that on the proposed dataset, our network achieved an mAP of 0.935, and an mAP of 0.744 on a 0.5 threshold. In terms of network size, our network exhibits a compact parameter count of 11,550,110, comparable to YOLOV8S. This designates an optimized network that achieves better performance without requiring excessive computational resources or sacrificing efficiency. Additionally, our network achieved an admirable FPS of 126.00, certifying real-time processing abilities appropriate for time applications. The high FPS empowers rapid inference, allowing the network to process even video streams efficiently. Overall, the outcomes highlight the superior performance and computational efficiency of our proposed FD network compared to existing versions.

5. Conclusion

This study presented significant advancements toward effective fall detection (FD) systems, especially for risky individuals such as impaired children, the elderly, and those with neurological problems. Through the execution of incremental improvements, we enhanced the capability of the proposed network to accurately detect falls in difficult and immediate situations. The incorporation of an extensive dataset, including more than 10,500 samples, provides a solid foundation for the training FD network. The data comprises a wide range of environmental conditions, to ensure the well generalized performance in real-life situations. In addition, our enhancements including the integration of a focus module in the backbone and a attention over attention mechanism in the neck module, significantly improve the features and the comprehension of spatial and channel contexts. The empirical analysis of our proposed network shows its higher performance in comparison to state-of-the-art methods. This is demonstrated by rigorous benchmarking and qualitative validation performed in different scenarios. The network's outstanding precision and computing speed indicate its capacity for practical use in assisting at-risk persons in their daily lives. Our future plans include examining the use of multi-modal sensor data by giving lightweight pipelines to improve understanding of fall incidents by considering multi-sensor data. In addition, we aim to further augment our dataset with a wider array of more challenging circumstances of different weather effects to ensure adaptability in practical settings.

CRediT authorship contribution statement

Habib Khan: Writing – original draft, Conceptualization, Visualization, Methodology. **Inam Ullah:** Writing – review & editing, Visualization, Data curation. **Mohammad Shabaz:** Methodology, Data curation, Formal analysis. **Muhammad Faizan Omer:** Data labeling, Data curation. **Muhammad Talha Usman:** Visualization, Formal analysis, Validation. **Mohammed Seghir Guellil:** Literature Search, Conceptualization. **JaKeoung Koo:** Writing – review & editing, Resources, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

Datasets with annotations and results will be available to the research community. The access/password will be granted on request for downloading: https://github.com/habib1402/Fall-Detection-DiverseFall10500.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. RS-2023-00240740).

References

[1] Falls, World Health Organization. https://www.who.int/news-room/fact-sheets/detail/falls, 1 October 2022.
[2] L. Ren, Y. Peng, Research of fall detection and fall prevention technologies: a systematic review, IEEE Access 7 (2019) 77702–77722, https://doi.org/10.1109/ACCESS.2019.2922708.
[3] L. Ma, M. Liu, N. Wang, L. Wang, Y. Yang, H. Wang, Room-level fall detection based on ultra-wideband (uwb) monostatic radar and convolutional long short-term memory (lstm), Sensors 20 (4) (2020) 1105.
[4] K. Wang, G. Zhan, W. Chen, 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, 2019, pp. 197–201.
[5] Z. Sheng-lan, Y. Yi-fan, G. Li-fu, W. Diao, Research and design of a fall detection system based on multi-axis sensor, in: Proceedings of the 4th International Conference on Intelligent Information Processing, 2019, pp. 303–309.
[6] P.V. Er, K.K. Tan, Wearable solution for robust fall detection, in: Assistive Technology for the Elderly, Elsevier, 2020, pp. 81–105.
[7] I. Charfi, J. Miteran, J. Dubois, M. Atri, R. Tourki, Definition and performance evaluation of a robust svm based fall detection solution, in: 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, IEEE, 2012, pp. 218–224.
[8] G. Mastorakis, D. Makris, Fall detection system using kinect's infrared sensor, J. Real-Time Image Proc. 9 (2014) 635–646.
[9] E. Alam, A. Sufian, P. Dutta, M. Leo, Vision-based human fall detection systems using deep learning: a review, Comput. Biol. Med. 146 (2022) 105626.
[10] Z. Zhang, C. Conly, V. Athitsos, Evaluating depth-based computer vision methods for fall detection under occlusions, in: Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8–10, 2014, Proceedings, Part II 10, Springer, 2014, pp. 196–207.
[11] A. Raza, M.H. Yousaf, S.A. Velastin, Human fall detection using yolo: A real-time and ai-on-the-edge perspective, in: 2022 12th International Conference on Pattern Recognition Systems (ICPRS), IEEE, 2022, pp. 1–6.
[12] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, C. Peñafort-Asturiano, Up-fall detection dataset: a multimodal approach, Sensors 19 (9) (2019) 1988.
[13] X. Wang, J. Ellul, G. Azzopardi, Elderly fall detection systems: a literature survey, Front. Robot. AI 7 (2020) 71.
[14] P. Qi, D. Chiaro, F. Piccialli, Fl-fd: federated learning-based fall detection with multimodal data fusion, Inform. Fusion 99 (2023) 101890.

[15] Y.M. Galvão, J. Ferreira, V.A. Albuquerque, P. Barros, B.J. Fernandes, A multimodal approach using deep learning for fall detection, Expert Syst. Appl. 168 (2021) 114226.
[16] F. Wahab, I. Ullah, A. Shah, R.A. Khan, A. Choi, M.S. Anwar, Design and implementation of real-time object detection system based on single-shoot detector and opencv, Front. Psychol. 13 (2022) 1039645.
[17] E. Lee, J.-S. Kim, D.K. Park, T. Whangbo, Yolo–mr: Meta-learning-based lesion detection algorithm for resolving data imbalance, in: IEEE Access, 2024.
[18] J. An, Route positioning system for campus shuttle bus service using a single camera, Electronics 13 (11) (2024) 2004.
[19] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.
[20] F. Lezzar, D. Benmerzoug, I. Kitouni, Camera-based fall detection system for the elderly with occlusion recognition, Appl. Med. Inform. 42 (3) (2020) 169–179.
[21] L. Killian, M. Julien, B. Kevin, L. Maxime, B. Carolina, C. Mélanie, B. Nathalie, G. Sylvain, G. Sebastien, 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, 2021, pp. 7–12.
[22] D. Zhao, T. Song, J. Gao, D. Li, Y. Niu, Yolo-fall: a novel convolutional neural network model for fall detection in open spaces, in: IEEE Access, 2024.
[23] Y. Ke, Y. Yao, Z. Xie, H. Xie, H. Lin, C. Dong, 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), 2023, pp. 0942–0949, https://doi.org/10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361490.
[24] D. Zhao, T. Song, J. Gao, D. Li, Y. Niu, Yolo-fall: a novel convolutional neural network model for fall detection in open spaces, IEEE Access 12 (2024) 26137–26149, https://doi.org/10.1109/ACCESS.2024.3362958.
[25] T. Chen, Z. Ding, B. Li, Elderly fall detection based on improved yolov5s network, IEEE Access 10 (2022) 91273–91282.
[26] C.-L. Liu, C.-H. Lee, P.-M. Lin, A fall detection system using k-nearest neighbor classifier, Expert Syst. Appl. 37 (10) (2010) 7174–7181.
[27] F. Kausar, M. Mesbah, W. Iqbal, A. Ahmad, I. Sayyed, Fall detection in the elderly using different machine learning algorithms with optimal window size, Mobile Netw. Appl. (2023) 1–11.
[28] B. Kwolek, M. Kepski, Improving fall detection by the use of depth sensor and accelerometer, Neurocomputing 168 (2015) 637–645.
[29] D. Yacchirema, J.S. De Puga, C. Palau, M. Esteve, Fall detection system for elderly people using iot and big data, Procedia Comp. Sci. 130 (2018) 603–610.
[30] E.E. Geertsema, G.H. Visser, M.A. Viergever, S.N. Kalitzin, Automated remote fall detection using impact features from video and audio, J. Biomech. 88 (2019) 25–32.
[31] M. Saleh, R.L.B. Jeannès, Elderly fall detection using wearable sensors: a low cost highly accurate algorithm, IEEE Sensors J. 19 (8) (2019) 3156–3164.
[32] O. Seredin, A. Kopylov, S.-C. Huang, D. Rodionov, A skeleton features-based fall detection using microsoft kinect v2 with one class-classifier outlier removal, Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 42 (2019) 189–195.
[33] L. Chen, R. Li, H. Zhang, L. Tian, N. Chen, Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch, Measurement 140 (2019) 215–226.
[34] I. Chandra, N. Sivakumar, C.B. Gokulnath, P. Parthasarathy, Iot based fall detection and ambient assisted system for the elderly, Clust. Comput. 22 (2019) 2517–2525.
[35] J.-S. Lee, H.-H. Tseng, Development of an enhanced threshold-based fall detection system using smartphones with built-in accelerometers, IEEE Sensors J. 19 (18) (2019) 8293–8302.
[36] C.M. Lee, J. Park, S. Park, C.H. Kim, Fall-detection algorithm using plantar pressure and acceleration data, Int. J. Precis. Eng. Manuf. 21 (2020) 725–737.
[37] C. Iuga, P. Drăgan, L. Buşoniu, Fall monitoring and detection for at-risk persons using a uav, IFAC-PapersOnLine 51 (10) (2018) 199–204.
[38] E.E. Geertsema, G.H. Visser, M.A. Viergever, S.N. Kalitzin, Automated remote fall detection using impact features from video and audio, J. Biomech. 88 (2019) 25–32.
[39] G.K. Hader, M.M. Ben Ismail, O. Bchir, Automatic fall detection using region-based convolutional neural network, Int. J. Inj. Control Saf. Promot. 27 (4) (2020) 546–557.
[40] Y. Ke, Y. Yao, Z. Xie, H. Xie, H. Lin, C. Dong, 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), IEEE, 2023, pp. 0942–0949.
[41] M.M. Islam, O. Tayan, M.R. Islam, M.S. Islam, S. Nooruddin, M.N. Kabir, M. R. Islam, Deep learning based systems developed for fall detection: a review, IEEE Access 8 (2020) 166117–166137.
[42] P. Wu, H. Li, N. Zeng, F. Li, Fmd-yolo: An efficient face mask detection method for covid-19 prevention and control in public, Image Vis. Comput. 117 (2022) 104341.
[43] K. Tong, Y. Wu, Deep learning-based detection from the perspective of small or tiny objects: a survey, Image Vis. Comput. 123 (2022) 104471.
[44] Z. Qu, L.-Y. Gao, S.-Y. Wang, H.-N. Yin, T.-M. Yi, An improved yolov5 method for large objects detection with multi-scale feature cross-layer fusion network, Image Vis. Comput. 125 (2022) 104518.
[45] J.C.E. Guerrero, E.M. España, M.M. Añasco, J.E.P. Lopera, Dataset for human fall recognition in an uncontrolled environment, Data Brief 45 (2022) 108610.
[46] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, J. Rousseau, Multiple cameras fall dataset, DIRO-Université de Montréal, Tech. Rep. 1350 (2010) 24.
[47] I. Charfi, J. Miteran, J. Dubois, M. Atri, R. Tourki, Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and adaboost-based classification, J. Electron. Imag. 22 (4) (2013) 041106.
[48] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji, Y. Li, Depth-based human fall detection via shape features and improved extreme learning machine, IEEE J. Biomed. Health Inform. 18 (6) (2014) 1915–1922.
[49] B. Kwolek, M. Kepski, Human fall detection on embedded platform using depth maps and wireless accelerometer, Comput. Methods Prog. Biomed. 117 (3) (2014) 489–501.
[50] M. Aslan, Y. Akbulut, A. Şengür, M.C. Ince, Skeleton based efficient fall detection, J. Fac. Eng. Archit. Gazi Univ. 32 (4) (2017) 1025–1034.
[51] K. Adhikari, H. Bouchachia, H. Nait-Charif, Activity recognition for indoor fall detection using convolutional neural network, in: 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), IEEE, 2017, pp. 81–84.
[52] D. Lee, J. Cho, Automatic object detection algorithm-based braille image generation system for the recognition of real-life obstacles for visually impaired people, Sensors 22 (4) (2022) 1601.
[53] S. Ahmad, J.-S. Kim, D.K. Park, T. Whangbo, Automated Detection of Gastric Lesions in Endoscopic Images by Leveraging Attention-Based yolov7, IEEE Access, 2023.
[54] M. Hussain, Yolov1 to v8: unveiling each variant–a comprehensive review of yolo, IEEE Access 12 (2024) 42816–42833.
[55] M. Hussain, Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection, Machines 11 (7) (2023) 677.
[56] X. Dong, S. Yan, C. Duan, A lightweight vehicles detection network model based on yolov5, Eng. Appl. Artif. Intell. 113 (2022) 104914.
[57] S.N. Saydirasulovich, M. Mukhiddinov, O. Djuraev, A. Abdusalomov, Y.-I. Cho, An improved wildfire smoke detection based on yolov8 and uav images, Sensors 23 (20) (2023) 8374.
[58] N.U.A. Tahir, Z. Long, Z. Zhang, M. Asim, M. ELAffendi, Pvswin-yolov8s: Uav-based pedestrian and vehicle detection for traffic management in smart cities using improved yolov8, Drones 8 (3) (2024) 84.
[59] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, 2021 arXiv preprint arXiv:2107.08430.
[60] A. Song, Z. Zhao, Q. Xiong, J. Guo, Lightweight the focus module in yolov5 by dilated convolution, in: 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), IEEE, 2022, pp. 111–114.
[61] M. Munsif, S.U. Khan, N. Khan, S.W. Baik, Attention-based deep learning framework for action recognition in a dark environment, Hum. Centric Comput. Inf. Sci 14 (2024) 1–22.
[62] H. Yar, Z.A. Khan, I. Rida, W. Ullah, M.J. Kim, S.W. Baik, An efficient deep learning architecture for effective fire detection in smart surveillance, Image Vis. Comput. 145 (2024) 104989.