applied sciences

MDPI

Article

Bridging the Appearance Domain Gap in Elderly Posture Recognition with YOLOv9

Andrés Bustamante [1], Lidia M. Belmonte [1,2], Rafael Morales [1,2], António Pereira [3,4] and Antonio Fernández-Caballero [1,5,6,*]

1 Instituto de Investigación en Informática de Albacete, 02071 Albacete, Spain; christian.bustamante@alu.uclm.es (A.B.); lidiamaria.belmonte@uclm.es (L.M.B.); rafael.morales@uclm.es (R.M.)
2 Departamento de Ingeniería Eléctrica, Electrónica, Automática y Comunicaciones, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
3 Computer Science and Communications Research Centre, Polytechnic Institute of Leiria, School of Technology and Management, 2411-901 Leiria, Portugal; apereira@ipleiria.pt
4 INOV INESC INOVAÇÃO, Institute of New Technologies—Leiria Office, 2411-901 Leiria, Portugal
5 Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
6 Biomedical Research Networking Centre in Mental Health (CIBERSAM), 28016 Madrid, Spain
* Correspondence: antonio.fdez@uclm.es; Tel.: +34-926-052-959

Abstract: Accurate posture detection of elderly people is crucial to improve monitoring and provide timely alerts in homes and elderly care facilities. Human posture recognition is experiencing a great leap in performance with the incorporation of deep neural networks (DNNs) such as YOLOv9. Unfortunately, DNNs require large amounts of annotated data for training, which can be addressed by using virtual reality images. This paper investigates how to address the appearance domain that lies between synthetic and natural images. Therefore, four experiments (VIRTUAL–VIRTUAL; HYBRID–VIRTUAL; VIRTUAL–REAL; and HYBRID–REAL) were designed to assess the feasibility of recognising the postures of virtual or real elderly people after training with virtual and real images of elderly people. The results show that YOLOv9 achieves the most outstanding accuracy of 98.41% in detecting and discriminating between *standing*, *sitting*, and *lying* postures after training on a large number of virtual images complemented by a much smaller number of real images when testing on real images.

Keywords: posture recognition; elderly care; appearance domain gap; virtual reality; YOLOv9

check for updates

Citation: Bustamante, A.; Belmonte, L.M.; Morales, R.; Pereira, A.; Fernández-Caballero, A. Bridging the Appearance Domain Gap in Elderly Posture Recognition with YOLOv9. *Appl. Sci.* **2024**, *14*, 9695. https://doi.org/10.3390/app14219695

Academic Editor: Andrea Prati

Received: 21 August 2024
Revised: 24 September 2024
Accepted: 22 October 2024
Published: 23 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The growing elderly population presents significant challenges to health and care systems, with falls being a major concern due to their potential to cause serious injury and loss of independence [1–3]. Traditionally, monitoring of the elderly in homes and care centres has relied heavily on human staff, sometimes limiting the capacity for rapid and continuous responses [2]. Surveillance by cameras equipped with artificial intelligence holds the promise of more effective monitoring and automatic alerting [4–7]. In particular, the ultimate aim of this study is to monitor the elderly through accurate identification of three key postures: *standing* , *sitting*, and *lying* .

The YOLO (You Only Look Once) family of models, renowned for their accuracy and efficiency in real-time object detection, have been successfully adapted for use in monitoring the elderly. We believe that YOLO's latest iteration, YOLOv9, is a very good solution for pushing the boundaries of real-time human activity detection because it can achieve high accuracy and speed. In fact, the training regime of YOLOv9 incorporates rigorous augmentation techniques and hyperparameter tuning to optimise performance. It also has the ability to learn complex feature representations and is efficient at handling large datasets.

However, to date, annotated datasets do not provide sufficient analysis of human activities, postures, or diversity of contexts [8]. In addition, privacy, legal, security, and ethical issues may limit the ability to collect sufficient human data. The use of human images in real-world scenarios is sometimes difficult due to privacy concerns. Given these issues, a popular approach is to feed the deep learning network with advanced forms of data, such as human skeletal joint trajectories, depth maps, or inertial data, instead of providing the network with raw data [9–12]. Another popular approach is using Internet of Things (IoT) wireless sensing, like mobile sensors on phones or watches [13,14].

Another alternative to real-world data that alleviates some of these problems is synthetic data [8]. Fortunately, the field of human posture recognition is experiencing a great leap in performance and capability due to the incorporation of deep neural networks [15–17]. Unfortunately, they require large amounts of annotated data for most applications. Acquiring adequate amounts of data is a tedious and expensive task. A powerful and cost-effective solution is to use synthetic data with automatically generated annotations [18]. In addition, domain adaptation is an elegant way to handle the lack of these datasets [19–21]. Finally, the use of synthetic images and/or videos generated by virtual reality software is becoming an attractive solution for training machine learning models in various application domains [22]. In particular, when focusing on human posture recognition, it is certainly important to have sufficiently large and varied image sets so that the models have a greater ability to generalise [23].

Recent studies have shown that synthetic images derived from simple 3D models can be used to train deep learning models to perform cross-domain synthetic-to-real visual recognition tasks [24,25]. However, the performance of such cross-domain models is degraded due to the appearance domain gap between the real and synthetic images [26,27]. In general, the appearance domain gap refers to visual differences between data from two domains caused by factors such as lighting, weather, time of day, and image resolution, among others. These differences can affect the performance of machine learning models, especially in computer vision tasks. Therefore, this work focuses on training the YOLOv9 model with synthetic and real videos to investigate the feasibility of later monitoring real elderly people and accurately identifying the three postures. Unity (https://unity.com/, accessed on 15 May 2024) was chosen for its powerful and flexible 3D simulation capabilities, allowing us to create realistic environments and scenarios to generate synthetic data [28]. Roboflow (https://roboflow.com/, accessed on 15 May 2024) was used for its robust data annotation and augmentation processes, which facilitate efficient training and validation of the model.

Although posture detection is an old topic, there is still room to study how to tackle the appearance domain gap. So far, the appearance domain gap between synthetic and natural images has been addressed using generative adversarial networks [18], association knowledge learning [19], and mainly transfer learning [29,30]. An inspiration for our work is a contribution from a paper that mixes training on large synthetic datasets with small real datasets for object recognition performance [31]. The novelty of the approach can be summarised as follows:

- By using Unity for realistic simulations and Roboflow for comprehensive data preparation, a new standard for automated elderly posture recognition is proposed.
- By using synthetic and real videos for YOLOv9 training, the efficiency of elderly posture detection in real life is enhanced, and the need for constant human supervision is reduced.
- The proposed method allows using a combination of a small number of real images with a much larger number of virtual images to train YOLOv9 for human posture detection in real-world images.

2. Literature Review

The use of YOLO is in line with recent research highlighting the effectiveness of certain models in posture detection, underlining their relevance and applicability in healthcare for

the elderly [1,6,32]. The first is an automated image-based fall detection system that can be used to monitor geriatric behaviour using the YOLOv3 algorithm [1], where fall events are identified and alerted as soon as they are detected. To meet the requirements of low cost, high accuracy, and real-time computation, the YOLOv3 object detection algorithm was combined with the IFADS fall detection method [33]. In addition, an improved YOLOv5s algorithm was recently proposed [6]. The work aims to improve the efficiency and accuracy of lightweight fall detection. Another fall detection system used the YOLOv8 algorithm [32]. The YOLOv8 algorithm helps detect whether a single person captured by surveillance cameras falls within an acceptable time frame.

Other advanced deep learning models have been used in elderly healthcare applications. For example, some works have focused on developing methods for elderly healthcare to identify human activity and achieve fall detection [34–37]. First, a mobile-enabled fall detection framework was presented in [34]. The framework is based on a hybrid deep learning model that combines a convolutional neural network (CNN) to extract the local representative features from the accelerometer sensor embedded in smartphones with a long short-term memory (LSTM) network to learn the dependencies between the features inferred from the sensor data. In addition, a fall detection method for elderly people was presented in [35]. It is based on an AlexNet CNN to produce a wearable IoT system that can detect falls. Data are collected from sensor devices placed on the subject's body in six different positions for feature extraction and sensor data analysis. Multi-linear principal component analysis is then applied to reduce the dimensionality of the derived features. Another fall detection method for elderly people was developed using a three-dimensional CNN as a suitable classifier for motion detection with a low-resolution infrared array sensor [36]. An LSTM network with feature extraction was also implemented for comparison. Finally, a radar-based fall detection method for elderly people was described, which is based on time-frequency analysis by applying the short-time Fourier transform to each radar return signal [37]. The resulting spectrograms are converted into binary images, which are fed into the CNN for high-level feature learning.

Other applications using deep learning models are related to health assistance in daily life [38–40]. For example, a system based on deep neural networks for identifying the activity of elderly residents in care facilities was presented in [38]. Another paper focused on the identification and prediction of anomalous behaviour by studying different deep neural network methodologies, including CNN-LSTM and Autoencoder CNN-LSTM [39]. Finally, a methodology based on a deep neural network, a two-layer framework, and marker-based stigmergy for identifying the activities of lonely older people was described in [40].

3. Materials and Methods

3.1. Experimental Design

Four experiments were designed to assess the feasibility of recognising the postures of virtual or real elderly people after training with virtual and real images of elderly people. The first experiment (VIRTUAL–VIRTUAL) was designed to investigate the performance of the proposal using only virtual images for YOLOv9 training, validation, and testing. Then, the second experiment (HYBRID–VIRTUAL) aimed to use virtual and real images during training and only virtual images for testing. The third experiment (VIRTUAL–REAL) focused on training and validation with virtual images only and testing with real images. Finally, the fourth experiment (HYBRID–REAL) used a hybrid set of virtual and real images for training. The test set was the same as in the second experiment, i.e., real images.

3.2. Virtual and Real Environments

For this study, a virtual home environment was created using Unity, as shown on the left side of Figure 1. The virtual home included two bedrooms, a dining area, and a hallway near the kitchen. Cameras were strategically placed in each corner of both bedrooms, as shown on the right side of Figure 1. The cameras were tilted at a 15-degree angle to the floor to simulate the typical tilt angle of cameras used in real homes. Each camera was

used to generate a video from which a set of images was extracted, resulting in a total of 1125 images.



Figure 1. (**Left**) Unity virtual environment. (**Right**) Top view; camera positions.

In addition, a real home environment was used in which an elderly person was located, as shown at the top of Figure 2. The room was selected because it is common and has large accessories such as flower pots and plants on each side of each chair, thus creating a somewhat congested environment. In this case, a single camera was used, which was located right in front of the room, above the dining table. Using a tripod, the camera was positioned so that it reached the height of the ceiling, as in the case of the cameras in the virtual environment. In this case, a video was recorded, and 41 frames were extracted (see the bottom of Figure 2). The collected data were manually annotated to create a ground-truth dataset to further ensure accurate evaluation of the model's performance.



Figure 2. (**Top**) Real environment. (**Bottom**) Real environment dataset.

3.3. Methodology

A video was recorded in the virtual environment to train, validate, and test our human posture recognition proposal in YOLOv9. The video showed three elderly people walking, sitting, and falling (lying). For each video frame (image), the model accurately enclosed

each individual in a bounding box labelled with the detected posture and a confidence score. Note that the bounding box with the confidence score was automatically added to the frame by the YOLO Ultralytics library (https://docs.ultralytics.com/, accessed on 15 May 2024) during the inference process.

The images (virtual and real) were uploaded to the Roboflow platform, where annotations were made for the three postures (*standing*, *sitting*, and *lying*). Roboflow facilitated an efficient data annotation and augmentation process. The annotations created a dataset, which was then downloaded in YOLOv9 format for training on a local machine using Python 3.10.12, CUDA 12.1, and the yolov9t.pt model.

The dataset was divided into three subsets: 87% for training, 8% for validation, and 4% for testing. The training parameters are detailed in Table 1, and the augmentation parameters used to improve model performance are described in Table 2. The dataset contains three labels: ['lying', 'sitting', and 'standing'], each representing a posture to be identified and classified.

Table 1. Training parameters.

Parameters	Value
Resize	True
imgsz	640
epochs	100
total images	1433
training images	1254
validating images	120
testing images	59

Table 2. Augmentation parameters.

Parameter	Value
Outputs per training example	3
Hue	Between −15° and +15°
Saturation	Between −25% and +25%
Brightness	Between −15% and +15%
Exposure	Between −10% and +10%
Blur	Up to 2.5 pixels
Noise	Up to 0.1% of pixels

3.3.1. Hyperparameters and Optimisation Process

The choice of hyperparameters was based on the need for efficient computation without compromising detection accuracy (see Table 1). The image size was resized to 640 × 640 pixels, and the model was trained for 100 epochs. The Adam optimiser, with an initial learning rate of 0.001, was chosen for its ability to adaptively adjust the learning rate and handle sparse gradients. A cosine learning rate scheduler was used to gradually decrease the learning rate during training, resulting in better convergence stability. The batch size was set to 16 to ensure efficient GPU utilisation during training.

3.3.2. Data Augmentation

In order to improve the generalisation ability of the model and to avoid overfitting, several data augmentation techniques were applied during the training phase (see Table 2). These included random hue adjustments between −15° and +15°, saturation changes between −25% and +25%, brightness changes between −15% and +15%, and exposure adjustments between −10% and +10%. Random noise affecting up to 0.1% of the pixels was also added to simulate real-world image imperfections. Greyscale transformations were applied to 15% of the images, and a blur of up to 2.5 pixels was introduced to make the model robust to varying image quality.

To ensure the robustness of the training process, the following steps were taken:

1. Data collection and annotation: The images extracted from the videos recorded by the six cameras in the virtual environment (used in the four experiments) and the images from the real-world video (used in the HYBRID–VIRTUAL and HYBRID–REAL experiments) were uploaded to Roboflow. The proportion of real data in the HYBRID dataset was 0.036. Each image was annotated for the three postures.
2. Data augmentation: Augmentation techniques, such as rotation, flipping, and scaling, were applied to the dataset to improve model generalisation. The specific augmentation parameters are listed in Table 2.
3. Model training: The YOLOv9 model was trained using the annotated and augmented dataset. Training was performed on a machine equipped with CUDA 11.8 to take advantage of GPU acceleration for faster training.

The methodology ensures that the model is well equipped to accurately detect and classify the three postures in different scenarios within the virtual environment. The remaining question is whether the application is consistent when tested in the real world.

3.4. Performance Evaluation

The performance metrics included precision, accuracy, recall, and F1-score for each posture category (*lying*, *sitting*, and *standing*), along with their respective averages. These metrics are defined as follows:

- Precision: The proportion of true positive detections out of the total number of detections for each posture.
- Accuracy: The overall correctness of the model's predictions, calculated as the ratio of correctly predicted instances to the total number of instances.
- Recall: The proportion of true positive detections among the actual instances for each posture.
- F1-score: The harmonic mean of precision and recall, providing a balance between the two metrics.

Two more metrics were introduced to ensure that the YOLOv9 model could accurately identify and localise objects within an image:

- mAP50: The mean average precision when evaluated at a single intersection over union (IoU) threshold of 0.50. This means that a predicted bounding box is considered a successful detection if it overlaps with the ground truth by at least 50%. This metric provides an indication of the model's ability to detect objects with a reasonable degree of accuracy.
- mAP50-95: This provides a more comprehensive evaluation by averaging precision across multiple IoU thresholds, ranging from 0.50 to 0.95 in increments of 0.05. This metric is more stringent, as it requires the model to perform well at varying levels of overlap, giving a better assessment of its overall detection performance.

4. Results

4.1. Training Results

In this paper, we compared the performance of the YOLOv9 model trained on two different datasets: a pure virtual dataset and a hybrid dataset. The results of the performance during the training phases are summarised in Figure 3.

Figure 3 shows that the trained YOLOv9 model was able to accurately detect and classify the postures of older people in both a virtual and a hybrid home environment. The model consistently identified the three postures—*standing* , *sitting*, and *lying* —with high confidence. The model showed consistent performance with similarly high precision, recall, mAP50, and mAP50-95 values for both environments.

Notably, the confidence values were mostly between 0.92 and 0.95, with a maximum of 0.97 and a minimum of 0.85.

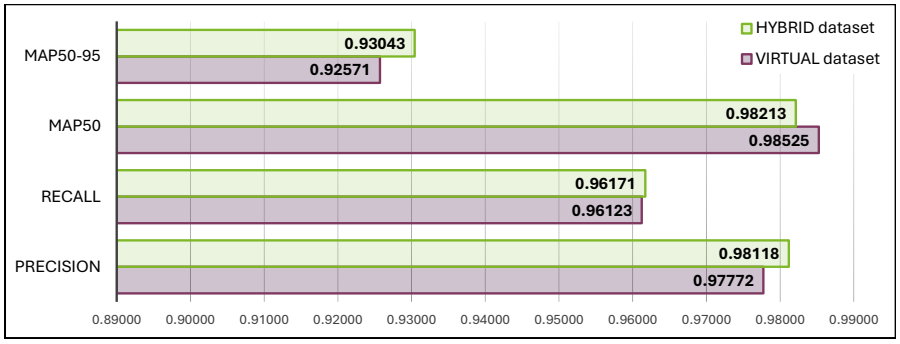


Figure 3. Comparison of training results with virtual and hybrid datasets.

4.2. Testing Results for the Virtual Environment

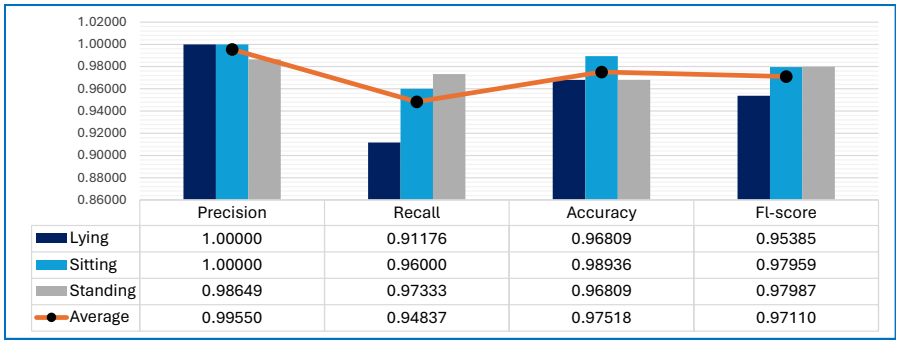
Figure 4 shows an example of posture detection in the virtual environment.



Figure 4. Example of results in a virtual environment with confidence scores.

4.2.1. Testing Results for VIRTUAL–VIRTUAL Experiment

The results of the virtual environment evaluation using the virtual dataset are presented in Figure 5.



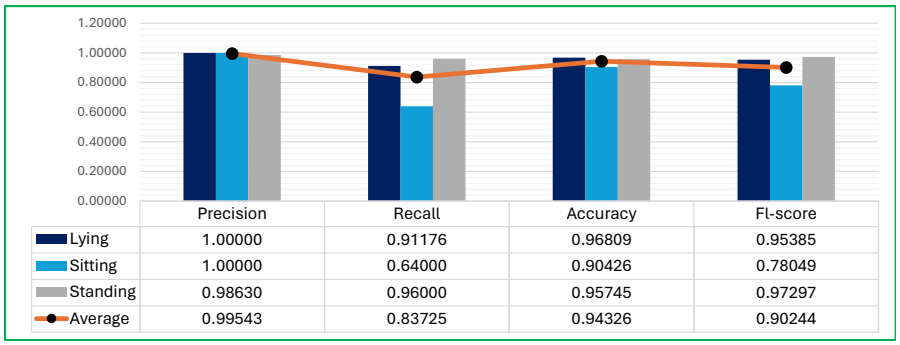
	Precision	Recall	Accuracy	Fl-score
Lying	1.00000	0.91176	0.96809	0.95385
Sitting	1.00000	0.96000	0.98936	0.97959
Standing	0.98649	0.97333	0.96809	0.97987
Average	0.99550	0.94837	0.97518	0.97110

Figure 5. Performance metrics for the VIRTUAL–VIRTUAL experiment.

The figure presents the performance metrics for the VIRTUAL–VIRTUAL experiment, including precision, recall, accuracy, and F1-score, across the three postures: *lying* , *sitting* , and *standing* . The precision metric is notably high, with perfect scores for *lying* and *sitting* and a near-perfect score for *standing* , resulting in an overall average of 0.99550. The recall values are slightly lower but still excellent, with the *sitting* posture achieving 0.96000 and the *lying* posture achieving 0.91176, bringing the average recall to 0.94837. The accuracy is outstanding and consistent across the postures, all above 0.96, leading to an average accuracy of 0.97518. The F1-score follows a similar pattern, with an overall average of 0.97110, indicating balanced performance between precision and recall.

4.2.2. Testing Results for HYBRID–VIRTUAL Experiment

Now, the results of the virtual environment evaluation using the hybrid dataset are presented in Figure 6.



	Precision	Recall	Accuracy	Fl-score
Lying	1.00000	0.91176	0.96809	0.95385
Sitting	1.00000	0.64000	0.90426	0.78049
Standing	0.98630	0.96000	0.95745	0.97297
Average	0.99543	0.83725	0.94326	0.90244

Figure 6. Performance metrics for the HYBRID–VIRTUAL experiment.

This new figure, detailing the performance metrics for the HYBRID–VIRTUAL experiment, shows similar precision scores to the VIRTUAL–VIRTUAL experiment. The precision remains high for *lying* and *sitting* , with perfect scores, and slightly lower for *standing* at 0.98630, leading to an average of 0.99543. However, the recall for *sitting* drops significantly to 0.64000, which lowers the overall average recall to 0.83725. The accuracy metric also reflects a slight decrease from the previous experiment, with *sitting* achieving 0.904626, whereas *lying* and *standing* maintain very high accuracy levels. This results in an overall average of 0.94326 (0.03 lower than for the previous experiment). The F1-score mirrors these trends, with the overall average for *standing* at 0.90244 (0.07 lower than for the VIRTUAL–VIRTUAL experiment).

4.3. Testing Results for Real Images

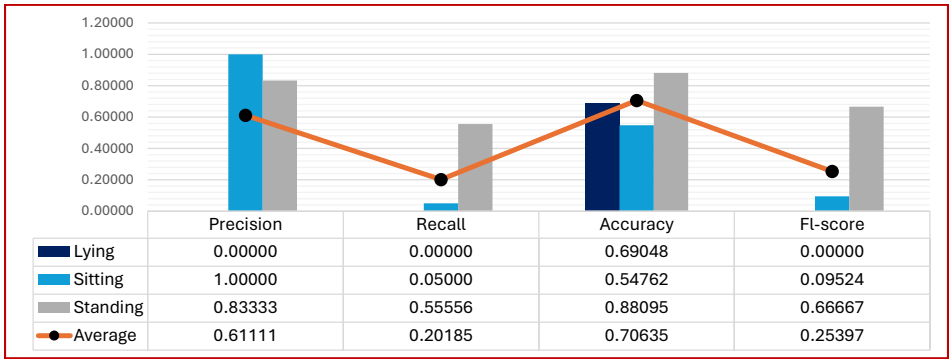
Figure 7 shows an example of posture detection in a real-world environment.



Figure 7. Example of results in a real environment with confidence scores.

4.3.1. Testing Results for VIRTUAL–REAL Experiment

The results of the real environment evaluation using the virtual dataset are presented in Figure 8.



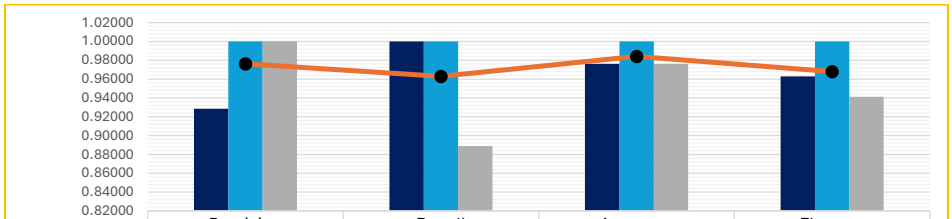
	Precision	Recall	Accuracy	Fl-score
Lying	0.00000	0.00000	0.69048	0.00000
Sitting	1.00000	0.05000	0.54762	0.09524
Standing	0.83333	0.55556	0.88095	0.66667
Average	0.61111	0.20185	0.70635	0.25397

Figure 8. Performance metrics for the VIRTUAL–REAL experiment.

The VIRTUAL–REAL experiment shows significant variability in performance across the postures. The precision for *lying* is 0.00000, indicating a complete failure to correctly identify this posture, while *sitting* achieves a perfect score, and *standing* is moderately successful with a precision of 0.83333, leading to a low overall average precision of 0.61111. The recall values for *lying* and *sitting* are also low, particularly for *lying* at 0.00000, resulting in an unacceptable average recall of 0.20185. The accuracy is relatively higher, with *lying* at 0.69048 and *standing* at 0.88095, culminating in an average accuracy of 0.70635. The F1-score reflects the poor performance in *lying* , with an overall catastrophic average of 0.25397.

4.3.2. Testing Results for HYBRID–REAL Experiment

The results of the real environment evaluation using the hybrid dataset are presented in Figure 9.



	Precision	Recall	Accuracy	Fl-score
Lying	0.92857	1.00000	0.97619	0.96296
Sitting	1.00000	1.00000	1.00000	1.00000
Standing	1.00000	0.88889	0.97619	0.94118
Average	0.97619	0.96296	0.98413	0.96805

Figure 9. Performance metrics for the HYBRID–REAL experiment.

In the HYBRID–REAL experiment, the precision is strong again across all postures, with *lying* at 0.92857, and both *sitting* and *standing* achieving perfect scores. This contributes to a high average precision of 0.97619. The recall is similarly strong, with all postures scoring high and an overall average recall of 0.96296. The accuracy metric remains consistently high across all postures, leading to an average accuracy of 0.98413. The F1-score also demonstrates balanced performance, with an average of 0.96805, reflecting again the strong alignment between precision and recall.

4.4. Real-Time Performance Evaluation

To evaluate the real-time performance of the proposed YOLOv9-based human posture recognition model, video frames captured from a webcam were processed on a system equipped with a Ryzen 7 7950X CPU (16 cores, 32 threads), 128 GB of RAM, and an NVIDIA

RTX 4080 GPU with 16 GB of VRAM. The inference was performed over a period of 1 minute, and key performance metrics were measured, including pre-processing time, inference time, post-processing time, and system resource usage (CPU, RAM, and GPU).

The results, as presented in Table 3, show that the average inference time per frame was 30.45 ms, with a minimum of 11.50 ms and a maximum of 72.58 ms. The pre-processing and post-processing times were minimal, averaging 1.93 ms and 1.63 ms, respectively. Resource usage was consistently low, with an average CPU utilisation of 0.77%, RAM utilisation of 14.24%, and GPU utilisation of 6.57%. GPU memory usage remained stable at an average of 2546.79 MB, indicating that the system maintained efficient memory management during the task.

Table 3. Real-time performance metrics for YOLOv9-based posture detection.

Metric	Min	Max	Avg
Pre-processing time (ms)	0.00	5.00	1.93
Inference time (ms)	11.5	72.58	30.45
Post-processing time (ms)	0.00	32.59	1.63
CPU usage (%)	0.30	2.20	0.77
RAM usage (%)	13.80	14.40	14.24
GPU usage (%)	0.00	21.00	6.57
GPU memory usage (MB)	2079.00	2585.00	2546.79

Based on the results, the system demonstrates strong performance in real-time posture recognition using YOLOv9. The average inference time of 30.45 ms per frame indicates that the model is capable of processing more than 30 frames per second (fps), which is suitable for real-time applications. The low CPU usage (average of 0.77%) and moderate GPU usage (average of 6.57%) suggest that the system uses the hardware resources efficiently and without significant load. In addition, the stability of GPU memory usage at around 2.5 GB out of 16 GB of VRAM shows effective memory management. Overall, these metrics suggest that the system is well optimised for real-time performance, making it suitable for use in environments that require responsive posture detection with minimal computational overhead.

4.5. Testing Results per Dataset

The radial plot presented in Figure 10 illustrates a comprehensive comparison between the performance of using a pure virtual dataset and a hybrid dataset. The radial plot was constructed by first calculating the average of each metric (precision, recall, accuracy and F1-score) for both datasets across the two validation environments.

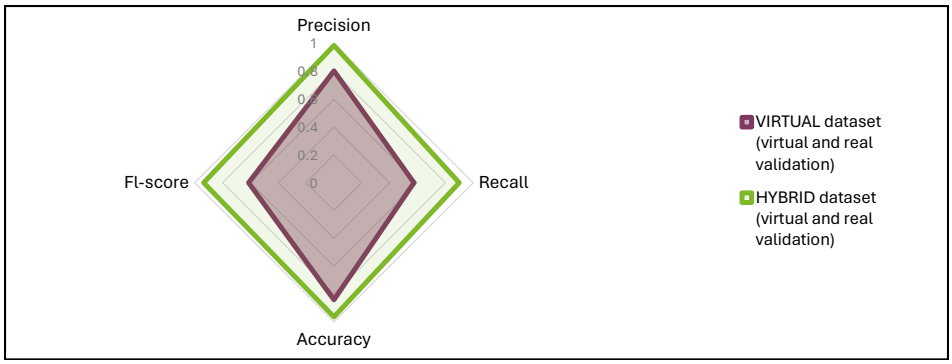


Figure 10. Radial comparison of results for the virtual dataset and the hybrid dataset.

The area under each line was filled to visually enhance the comparison. This visualisation helps in determining whether the incorporation of a small number of real images into the predominantly virtual dataset provides a significant improvement in model performance. The graph clearly shows the following:

1. The hybrid dataset shows a more balanced and generally higher performance across most metrics compared to the virtual dataset. This suggests that adding real images to the virtual dataset can enhance the model's ability to generalise better, particularly in a real-world validation environment.
2. While the virtual dataset performs exceptionally well in a controlled virtual environment, its performance drops significantly when validated in a real-world setting. The hybrid dataset, while slightly lower in virtual settings, maintains more consistent performance across different environments.

5. Discussion

Using YOLO for different human posture recognition tasks is not new. For instance, research with YOLOv5s [6] has highlighted improvements in fall detection, while YOLOv8 [32] has highlighted the potential of YOLOv8 in healthcare monitoring applications. The potential of YOLO models extends beyond fall detection. As suggested by studies of YOLOv4 and LSTM [41] and other advanced YOLO networks [42,43], these models show promise in detecting different postures and activities, potentially aiding broader healthcare applications. Research efforts have also been directed at optimising YOLO models for better performance. Novel approaches such as GL-YOLO-Lite [44] indicate continued improvements in model accuracy and efficiency. In addition, recent developments in human posture detection, such as CHP-YOLOF and CHP-YOLOX [45], have extended the capabilities of these systems to a wide range of human activities.

Our approach used YOLOv9 because it builds on the strengths of its predecessors and offers improvements in architecture and optimisation techniques. Although the core architectures of YOLOv8 and YOLOv9 are similar, YOLOv9 introduces refinements in the backbone and the head of the model to provide better feature extraction and more efficient handling of computational resources. These improvements are particularly relevant in scenarios that require high performance with minimal latency, such as real-time detection tasks. In addition, the adoption of YOLOv9 provided the opportunity to work with the latest developments in the YOLO framework, ensuring that the model benefited from the latest research and optimisations. In summary, while the specific task at hand could have been achieved with previous models such as YOLOv8, the decision to use YOLOv9 was driven by the desire to take advantage of the latest developments in the YOLO architecture, ensuring optimal performance and future-proofing the model.

Specifically, our proposal used the YOLOv9 model to efficiently detect three important postures for monitoring elderly people at home. The authors recognise that classifying three postures to study human behaviour may seem limited. For example, a recent study [46] looked at seven activities in the ADLF (Activities of Daily Living and Falls) dataset, including standing, sitting, lying, and falling. Another paper [9] that includes standing, sitting, lying, and falling also classified seven postures. However, the classification of a smaller number of human activities has also been described very recently [47]. In that paper, the four postures of standing, sitting, lying, and dangerous sitting were studied. In our case, working on three widely studied postures is only a first step, and we will certainly expand our approach to classify a greater number of postures. However, we would like to emphasise that the main objective of this work was to study the possibility of bridging the gap in the appearance domain between real and synthetic data using YOLOv9.

The current approach provided several important insights into the performance of the YOLOv9 model in different environments and with different datasets composed of synthetic and real images:

- The performance of the model in recognising postures in the virtual environment using the hybrid dataset showed a slight decrease in all metrics compared to using the virtual dataset. For example, the mean F1-score for the three postures was 0.97 when using the virtual dataset and 0.90 when using the hybrid dataset. This can be attributed to the gap in the appearance domain caused by the difference between the natural and synthetic image formation processes. A real image is the product of an image

acquisition process that inherently captures a variety of different natural phenomena, such as motion noise and material properties. On the other hand, a synthetic image is the result of a light transport simulation and rendering [18].

- The real environment recognition results using the hybrid dataset showed impressive improvements in all metrics compared to the virtual dataset. For example, the average F1-score was 0.25 for the VIRTUAL–REAL experiment and 0.97 for the HYBRID–REAL experiment, almost four times higher. Here, on the other hand, the appearance domain gap was largely bridged by including some real (natural) images in the virtual dataset in order to detect postures in the real world.

More concretely, when looking at the different metrics, Figure 11 helps in providing very meaningful information.

Across all experiments, the precision for the *sitting* posture consistently achieved a perfect score except in the VIRTUAL–REAL experiment, where the overall average precision was significantly lower. This indicates that *sitting* was generally well identified by the model, except when real-world data were introduced, as seen in the VIRTUAL–REAL experiment. The recall metric showed the most variability, especially in the VIRTUAL–REAL experiment, where the model failed to detect the *lying* posture entirely. This suggests that the model struggled with detecting certain postures when transitioning from a virtual to a real environment. The accuracy remained high in most experiments, particularly in the HYBRID–REAL experiment, indicating that combining virtual and real data led to more reliable performance across all postures. However, the significant drop in accuracy for the *sitting* posture in the HYBRID–VIRTUAL experiment (0.64000) highlights potential challenges when mixing datasets. The F1-score followed the trends observed for the recall. Behind the VIRTUAL–VIRTUAL case, the HYBRID–REAL experiment achieved the highest overall F1-score, suggesting that the integration of real data with virtual data significantly improved the model's performance. The lowest F1-score was observed in the VIRTUAL–REAL experiment, again highlighting the difficulties in generalising from virtual to real environments.

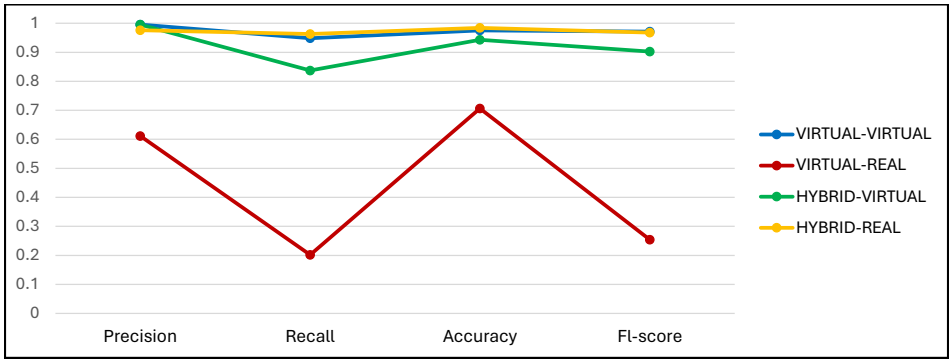


Figure 11. Comparison of results per metric of the 4 experiments.

Overall, the HYBRID–REAL experiment demonstrated the best overall performance across all metrics, suggesting that the combination of virtual and real data is the most effective strategy for enhancing experiences with real-world images. On the other hand, the VIRTUAL–REAL experiment showed the poorest performance, indicating that the model struggled the most when tested on real data without prior exposure to natural images. These comparisons emphasise the importance of incorporating real data during the training process, as seen in the improved metrics of the HYBRID–REAL experiment. Additionally, the variability in the recall and F1-score across different postures highlights the challenges in detecting certain postures consistently, especially when transitioning from virtual to real environments.

Thus, the results indicate that while the model performs exceptionally well in the virtual environment with both datasets, its performance in real-world scenarios is vastly im-

proved when trained with a hybrid dataset. This highlights the need to include real-world data in the training process to ensure the effectiveness of the model in practical applications.

6. Conclusions

This study has demonstrated the feasibility of using virtual environments to train posture recognition models for elderly people. The results show that while the YOLOv9 model trained solely on virtual data performs well in virtual environments, its real-world performance is severely limited due to the appearance gap problem. However, by including a small number of natural images in the training process (hybrid dataset), the performance of the model in real-world scenarios is significantly improved. According to a very recent survey [48], compared to our closest deep learning competitors in vision-based human activity recognition, our approach aligns well with their classification results in terms of accuracy.

Future work will focus on further investigating the proportion of real-world images to be used in the training set to improve the robustness of the YOLOv9 model. In addition, we plan to explore content domain adaptation techniques to better match the characteristics of virtual and real data, thereby improving the generalisability of the YOLOv9 model across different environments. Expanding the dataset and using more datasets to include a wider variety of postures and real-world scenarios will be crucial for developing a comprehensive posture recognition model. Finally, a comparison of our proposed model with different YOLO models, such as YOLOv8, YOLOv7, YOLOX, and YOLOR, will demonstrate which is the most suitable for human posture recognition among the YOLO family.

Author Contributions: Conceptualisation, R.M. and L.M.B.; methodology, A.B. and L.M.B.; software, A.B.; validation, R.M., A.P. and A.F.-C.; formal analysis, A.B. and L.M.B.; writing—original draft preparation, A.B. and L.M.B.; writing—review and editing, A.F.-C.; supervision, R.M.; funding acquisition, A.P. and A.F.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grant PID2020-115220RB-C21, funded by the Spanish MICIU/AEI/10.13039/501100011033 and "ERDF: A Way to Make Europe". This research was also supported by CIBERSAM, Instituto de Salud Carlos III, MICIU, and "ERDF: A Way to Make Europe". Additionally, grant 2022-GRIN-34436 was funded by Universidad de Castilla-La Mancha and "ERDF: A Way to Make Europe". This work was partially supported by national funds through the Portuguese Foundation for Science and Technology (FCT), I.P., under the project UIDB/04524/2020. It was also partially supported by Junta de Comunidades de Castilla-La Mancha/ESF (grant No. SBPLY/21/180501/000030).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DNN	Deep Neural Network
YOLO	You Only Look Once

References

1. Kavitha, A.; Hemalatha, B.; Abishek, K.; Harigokul, R. Fall Detection of Elderly Using YOLO. In *Proceedings of the ICT Systems and Sustainability*; Tuba, M., Akashe, S., Joshi, A., Eds.; Springer: Singapore, 2023; pp. 113–121.
2. Raghav, A.; Chaudhary, S. Elderly Patient Fall Detection Using Video Surveillance. In *Proceedings of the Computer Vision and Image Processing*; Raman, B., Murala, S., Chowdhury, A., Dhall, A., Goyal, P., Eds.; Springer: Cham, Switzerland, 2022; pp. 450–459.
3. Sokolova, M.V.; Serrano-Cuerda, J.; Castillo, J.C.; Fernández-Caballero, A. A fuzzy model for human fall detection in infrared video. *J. Intell. Fuzzy Syst.* **2013**, *24*, 215–228. [CrossRef]

4. Rojas-Albarracín, G.; Fernández-Caballero, A.; Pereira, A.; López, M.T. Heart Attack Detection Using Body Posture and Facial Expression of Pain. In *Proceedings of the Artificial Intelligence for Neuroscience and Emotional Systems*; Ferrández Vicente, J.M., Val Calvo, M., Adeli, H., Eds.; Springer: Cham, Switzerland, 2024; pp. 411–420.
5. Roda-Sanchez, L.; Garrido-Hidalgo, C.; García, A.; Teresa, O.; Fernández-Caballero, A. Comparison of RGB-D and IMU-based gesture recognition for human-robot interaction in remanufacturing. *Int. J. Adv. Manuf. Technol.* **2023**, *124*, 3099–3111. [CrossRef]
6. Wang, Y.; Chi, Z.; Liu, M.; Li, G.; Ding, S. High-Performance Lightweight Fall Detection with an Improved YOLOv5s Algorithm. *Machines* **2023**, *11*, 818. [CrossRef]
7. Bustamante, A.; Belmonte, L.M.; Pereira, A.; González, P.; Fernández-Caballero, A.; Morales, R. Vision-Based Human Posture Detection from a Virtual Home-Care Unmanned Aerial Vehicle. In *Proceedings of the Bio-Inspired Systems and Applications: From Robotics to Ambient Intelligence*; Ferrández Vicente, J.M., Álvarez-Sánchez, J.R., de la Paz López, F., Adeli, H., Eds.; Springer: Cham, Switzerland, 2022; pp. 482–491.
8. Ebadi, S.E.; Jhang, Y.C.; Zook, A.; Dhakad, S.; Crespi, A.; Parisi, P.; Borkman, S.; Hogins, J.; Ganguly, S. PeopleSansPeople: A Synthetic Data Generator for Human-Centric Computer Vision. *arXiv* **2022**, arXiv:2112.09290.
9. Yadav, S.K.; Tiwari, K.; Pandey, H.M.; Akbar, S.A. Skeleton-based human activity recognition using ConvLSTM and guided feature learning. *Soft Comput.* **2022**, *26*, 877–890. [CrossRef]
10. Usmani, A.; Siddiqui, N.; Islam, S. Skeleton joint trajectories based human activity recognition using deep RNN. *Multimed. Tools Appl.* **2023**, *82* , 46845–46869. [CrossRef]
11. Singh, R.; Khurana, R.; Kushwaha, A. Combining CNN streams of dynamic image and depth data for action recognition. *Multimed. Syst.* **2020**, *26*, 313–322. [CrossRef]
12. Dentamaro, V.; Gattulli, V.; Impedovo, D.; Manca, F. Human activity recognition with smartphone-integrated sensors: A survey. *Expert Syst. Appl.* **2024**, *246*, 123143. [CrossRef]
13. Hosseinzadeh, M.; Koohpayehzadeh, J.; Ghafour, M.Y.; Ahmed, A.M.; Asghari, P.; Souri, A.; Pourasghari, H.; Rezapour, A. An elderly health monitoring system based on biological and behavioral indicators in internet of things. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 5085–5095. [CrossRef]
14. Piao, X. Design of Health and Elderly Care Intelligent Monitoring System Based on IoT Wireless Sensing and Data Mining. *Mob. Netw. Appl.* **2024**, *29*, 153–167. [CrossRef]
15. Jiang, X.; Hu, Z.; Wang, S.; Zhang, Y. A Survey on Artificial Intelligence in Posture Recognition. *Comput. Model. Eng. Sci.* **2023**, *137*, 35–82. [CrossRef] [PubMed]
16. Lee, M.F.R.; Chen, Y.C.; Tsai, C.Y. Deep Learning-Based Human Body Posture Recognition and Tracking for Unmanned Aerial Vehicles. *Processes* **2022**, *10*, 2295. [CrossRef]
17. Ogundokun, R.O.; Maskeliūnas, R.; Damaševičius, R. Human Posture Detection Using Image Augmentation and Hyperparameter-Optimized Transfer Learning Algorithms. *Appl. Sci.* **2022**, *12*, 10156. [CrossRef]
18. Kviatkovsky, I.; Bhonker, N.; Medioni, G. From Real to Synthetic and Back: Synthesizing Training Data for Multi-Person Scene Understanding. *arXiv* **2020**, arXiv:2006.02110.
19. Liu, X.; Yoo, C.; Xing, F.; Oh, H.; El Fakhri, G.; Kang, J.W.; Woo, J. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Trans. Signal Inf. Process.* **2022**, *11*, e25. [CrossRef]
20. Himeur, Y.; Al-Maadeed, S.; Kheddar, H.; Al-Maadeed, N.; Abualsaud, K.; Mohamed, A.; Khattab, T. Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105698. [CrossRef]
21. Singhal, P.; Walambe, R.; Ramanna, S.; Kotecha, K. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access* **2023**, *11*, 6973–7020. [CrossRef]
22. Anvari, T.; Park, K.; Kim, G. Upper Body Pose Estimation Using Deep Learning for a Virtual Reality Avatar. *Appl. Sci.* **2023**, *13*, 2460. [CrossRef]
23. Romero, A.; Carvalho, P.; Côrte-Real, L.; Pereira, A. Synthesizing Human Activity for Data Generation. *J. Imaging* **2023**, *9*, 204. [CrossRef]
24. Reddy, A.V.; Shah, K.; Paul, W.; Mocharla, R.; Hoffman, J.; Katyal, K.D.; Manocha, D.; de Melo, C.M.; Chellappa, R. Synthetic-to-Real Domain Adaptation for Action Recognition: A Dataset and Baseline Performances. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation, London, UK, 29 May–2 June 2023; pp. 11374–11381. [CrossRef]
25. Acharya, D.; Tatli, C.J.; Khoshelham, K. Synthetic-real image domain adaptation for indoor camera pose regression using a 3D model. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 405–421. [CrossRef]
26. Liu, Y.; Wang, Z.; Zhou, X.; Zheng, L. A Study of Using Synthetic Data for Effective Association Knowledge Learning. *Mach. Intell. Res.* **2023**, *20*, 194–206. [CrossRef]
27. Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; Gong, B. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Los Alamitos, CA, USA, 2019; pp. 2100–2110. [CrossRef]
28. Bustamante, A.; Belmonte, L.M.; Morales, R.; Pereira, A.; Fernández-Caballero, A. Video Processing from a Virtual Unmanned Aerial Vehicle: Comparing Two Approaches to Using OpenCV in Unity. *Appl. Sci.* **2022**, *12*, 5958. [CrossRef]

29. Hennicke, L.; Adriano, C.M.; Giese, H.; Koehler, J.M.; Schott, L. Mind the Gap Between Synthetic and Real: Utilizing Transfer Learning to Probe the Boundaries of Stable Diffusion Generated Data. *arXiv* **2024**, arXiv:2405.03243.
30. Li, Y.; Dong, X.; Chen, C.; Li, J.; Wen, Y.; Spranger, M.; Lyu, L. Is Synthetic Image Useful for Transfer Learning? An Investigation into Data Generation, Volume, and Utilization. *arXiv* **2024**, arXiv:2403.19866.
31. Nowruzi, F.E.; Kapoor, P.; Kolhatkar, D.; Hassanat, F.A.; Laganière, R.; Rebut, J. How much real data do we actually need: Analyzing object detection performance using synthetic and real data. *arXiv* **2019**, arXiv:1907.07061.
32. Ng, P.H.; Mai, A.; Nguyen, H. Building an AI-Powered IoT App for Fall Detection Using Yolov8 Approach. In *Proceedings of the Intelligence of Things: Technologies and Applications*; Dao, N.N., Thinh, T.N., Nguyen, N.T., Eds.; Springer: Cham, Switzerland, 2023; pp. 65–74.
33. Lu, K.L.; Chu, E.T.H. An Image-Based Fall Detection System for the Elderly. *Appl. Sci.* **2018**, *8*, 1995. [CrossRef]
34. Hassan, M.M.; Gumaei, A.; Aloi, G.; Fortino, G.; Zhou, M. A Smartphone-Enabled Fall Detection Framework for Elderly People in Connected Home Healthcare. *IEEE Netw.* **2019**, *33*, 58–63. [CrossRef]
35. Alarifi, A.; Alwadain, A. Killer heuristic optimized convolution neural network-based fall detection with wearable IoT sensor devices. *Measurement* **2021**, *167*, 108258. [CrossRef]
36. Tateno, S.; Meng, F.; Qian, R.; Hachiya, Y. Privacy-Preserved Fall Detection Method with Three-Dimensional Convolutional Neural Network Using Low-Resolution Infrared Array Sensor. *Sensors* **2020**, *20*, 5957. [CrossRef]
37. Sadreazami, H.; Bolic, M.; Rajan, S. Contactless Fall Detection Using Time-Frequency Analysis and Convolutional Neural Networks. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6842–6851. [CrossRef]
38. Minvielle, L.; Audiffren, J. NurseNet: Monitoring Elderly Levels of Activity with a Piezoelectric Floor. *Sensors* **2019**, *19*, 3851. [CrossRef] [PubMed]
39. Zerkouk, M.; Chikhaoui, B. Spatio-Temporal Abnormal Behavior Prediction in Elderly Persons Using Deep Learning Models. *Sensors* **2020**, *20*, 2359. [CrossRef] [PubMed]
40. Xu, Z.; Wang, G.; Guo, X. Sensor-based activity recognition of solitary elderly via stigmergy and two-layer framework. *Eng. Appl. Artif. Intell.* **2020**, *95*, 103859. [CrossRef]
41. Chutimawattanakul, P.; Samanpiboon, P. Fall Detection for The Elderly using YOLOv4 and LSTM. In Proceedings of the 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Prachuap Khiri Khan, Thailand, 24–27 May 2022; pp. 1–5. [CrossRef]
42. Chen, T.; Ding, Z.; Li, B. Elderly Fall Detection Based on Improved YOLOv5s Network. *IEEE Access* **2022**, *10*, 91273–91282. [CrossRef]
43. Raza, A.; Yousaf, M.H.; Velastin, S.A. Human Fall Detection using YOLO: A Real-Time and AI-on-the-Edge Perspective. In Proceedings of the 2022 12th International Conference on Pattern Recognition Systems (ICPRS), Saint-Etienne, France, 7–10 June 2022; pp. 1–6. [CrossRef]
44. Dai, Y.; Liu, W. GL-YOLO-Lite: A Novel Lightweight Fallen Person Detection Model. *Entropy* **2023**, *25*, 587. [CrossRef]
45. Li, Y.; Wu, Y.; Chen, X.; Chen, H.; Kong, D.; Tang, H.; Li, S. Beyond Human Detection: A Benchmark for Detecting Common Human Posture. *Sensors* **2023**, *23*, 8061. [CrossRef]
46. Yadav, S.K.; Luthra, A.; Tiwari, K.; Pandey, H.M.; Akbar, S.A. ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling. *Knowl.-Based Syst.* **2022**, *239*, 107948. [CrossRef]
47. Guerra, B.M.V.; Ramat, S.; Beltrami, G.; Schmid, M. Recurrent Network Solutions for Human Posture Recognition Based on Kinect Skeletal Data. *Sensors* **2023**, *23*, 5260. [CrossRef]
48. Shafizadegan, F.; Naghsh-Nilchi, A.; Shabaninia, E. Multimodal vision-based human action recognition using deep learning: A review. *Artif. Intell. Rev.* **2024**, *54*, 178. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.