A Vision-based Human Posture Detection Approach for Smart Home Applications

Yangxia Shu[1]*, Lei Hu[2]
College of Big Data Science, Jiangxi Institute of Fashion Technology, Nanchang 330201, Jiangxi, China[1]
Information Technology Integration Innovation Center, Intelligent Research and Innovation Team for Clothing, Jiangxi Institute of Fashion Technology, Nanchang 330201, Jiangxi, China[1, 2]
Operation and Maintenance Section of Asset Department, Jiangxi Institute of Fashion Technology, Nanchang 330201, Jiangxi, China[2]

***Abstract*—Effective posture identification in smart home applications is a challenging topic for people to tackle in order to decrease the occurrence of improper postures. Vision-based posture identification has been used to construct a system for identifying people's postures. However, the system complexity, low accuracy rate, and slow identification speed of existing vision-based systems make them unsuitable for smart home applications. The goal of this project is to address these issues by creating a vision-based posture recognition system that can recognize human position and be used in smart home applications. The suggested method involves training and testing a You Only Look Once (YOLO) network to identify the postures. This Yolo-based approach is based on YOLOv5, which provides a high accuracy rate and satisfied speed in posture detection. Experimental results show the effectiveness of the developed system for posture recognition on smart home applications.**

***Keywords*—Posture identification; smart home applications; vision-based recognition; YOLO network; accuracy**

I. Introduction

Recognition of human body position has been extensively utilized in the fields of detection and rescue, intelligent monitoring, and other areas of computer vision as one of the most significant research paths. Its major objective is to study different human body areas, extract posture information, and eventually recognize human body position using computer vision technology [1].

Healthcare concerns are suddenly becoming more and more crucial as the number of senior individuals worldwide rises. Human motion capture technologies are necessary for older adults who live alone in order to address these problems. Seniors' health may also be tracked by observing their posture, and if high-risk postures, such as falling over, are noticed, a warning can be sent. These solutions will ease the strain on human resources while enhancing posture recognition effectiveness [2]. However, due to variables such as shifting viewing angles, human body occlusion, and appearance variations, determining the human posture with accuracy is a highly difficult process [1, 3].

In recent years, there has been remarkable progress in image processing, thanks to the advancements in deep learning, particularly convolutional neural networks (CNN). These networks draw inspiration from the hierarchical processing observed in the human visual cortex. By employing CNN, the process of feature extraction and classification has been revolutionized, as it allows the network to discern crucial features from the provided training data automatically. As a result, CNN has demonstrated remarkable success in accurate image processing [4]. A multi-layer neural network called CNN is mostly employed to scale and recognize displacement in two-dimensional visuals. The convolutional neural network's layers each represent a change. Common methods include convention and pooling transformation. Each transformation expresses different features from the input features as well processes the input data in a specific way. Each layer is made up of several two-dimensional planes containing the feature map that each layer has processed [5]. Each output characteristic is also an input feature; however, the value's computation process is the same. It is compatible with the generic neural network since it is the dot product of the weight and the input before the bias is imposed.

Within the realm of posture detection algorithms based on CNN, two primary types can be identified. The first category consists of two-stage detection algorithms, which involve a two-step process: locating the target and then recognizing it. Among these, the Region-Convolutional Neural Network (R-CNN) is a well-known traditional technique. However, it has shown poor performance and fails to meet real-time processing demands [6]. To address this limitation, subsequent advancements were made, giving rise to the Fast regions with CNN (Fast R-CNN) and faster regions with CNN (Faster R-CNN). Despite these improvements, they still do not fully satisfy real-time expectations [6]. The second category involves a one-stage detection technique, streamlining target localization and identification into a single action. Within this approach, examples like the single shot multi-box detector (SSD) series and the you only look once (YOLO) series are commonly cited as traditional instances of this methodology [7]. These methods have been developed to achieve faster and more efficient posture detection compared to the two-stage approaches.

The YOLO is regarded as one of the most effective and least time-consuming posture identification techniques. The YOLO model is used in several applications, including recognizing pedestrians and cars in traffic situations, monitoring livestock, aerial analysis, and even helping the visually handicapped identify faces. YOLO is an accurate posture detection method that combines a grid methodology

with the CNN architecture [8]. The YOLOv5 algorithm has been proposed as a modification of the YOLO algorithm version. On the basis of greater precision and a smaller model, YOLOv5's detection speed has significantly increased when compared to YOLOv3. The YOLOv5 technique has not yet found widespread application in the field of fall detection. Thus, this paper will enhance the model based on YOLOv5 research and apply it to the detection of senior fall behavior. The four network models of the target detection network based on YOLOv5 are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The three models of YOLOv5m, YOLOv5l, and YOLOv5x are the results of continual deepening and broadening based on YOLOv5s, which have the least depth and the smallest feature map width among them [7]. The YOLOv5 network is divided into the following four sections: the neck, the backbone, and the prediction.

The research contributions of this study are listed as follows:

1) The research contributes to the field of smart home applications by addressing the challenges of effective posture identification. It proposes a vision-based posture recognition system that overcomes the limitations of existing systems, such as complexity, low accuracy, and slow identification speed, thus making it suitable for integration into smart homes.

2) The study introduces a novel approach to posture recognition by leveraging the You Only Look Once (YOLO) network, specifically YOLOv5, which significantly improves accuracy and detection speed. This innovation enhances the viability of posture recognition systems for real-time monitoring and intervention in smart home environments.

3) Experimental results confirm the effectiveness of the developed YOLO-based posture recognition system in smart home applications. This research contribution showcases the practical applicability of the proposed method for reducing improper postures and enhancing the overall well-being of smart home residents.

The paper is organized as follows: Section II provides an introduction to the background of the study, offering essential context to the reader. In Section III, a comprehensive explanation of the methodology is presented, covering both the training and testing processes involved in human posture detection using the YOLO 5 network. Section IV delves into a detailed discussion and analysis of the obtained results, shedding light on the outcomes of the experiments and their implications. Finally, in Section V, the paper concludes with a summary of the findings and potential future directions for further research in this domain.

II. Related Works

In study [9], a technique for skeleton-based online HAR employing ST-GCN with a sliding window and majority voting approach using convolutional neural networks (STGCN-SWMV) has been developed in order to address the issue of online HAR for continuous flow skeleton data. Two online skeleton-based datasets called OAD and UOW have been used to assess this approach. By automatically learning geographical and temporal information, this method offers greater prediction power and generalizability. The goal of this work is to increase online recognition performance, and it is suggested that in order to do this, textual information about item appearance or human interaction may be integrated to provide more learning characteristics.

A deep convolutional neural network has been proposed as a method for recognizing human action from depth maps and posture data. To maximize feature extraction, three convolutional neural network channels and two action representations are combined. The trials demonstrate that employing three channels instead of one or using two channels alone produces superior outcomes. In order to classify human activities using one or two CNN channels at most for quick computations, the challenge of this work is to reduce the number of CNN channels [4].

To address these problems, the study in [10] demonstrates the Trajectory-weighted Deep Convolutional Rank-Pooling Descriptor (TDRD) for fall detection, which is resilient to surrounding settings and can successfully represent the dynamics of human motions in extended films. The SDUF all dataset had better results with TDRD, and the UR dataset and multi-camera datasets with SVM classifiers saw equivalent performance. The problem with the TDRD algorithm is that although it excels at detecting single falls, it struggles to do so in scenarios involving many people. Additionally, TDRD is a problem with characterizing prolonged static postures.

The author in [11] presented an innovative approach to human fall detection, leveraging the Fast Pose Estimation technique. The proposed method involves classifying data extracted from image frames using two models: Time-Distributed Convolutional Neural Network Long Short-Term Memory (TD-CNN-LSTM) and 1-Dimensional Convolutional Neural Network (1D-CNN). The results demonstrate impressive accuracy, making this technique a valuable addition to the realm of reliable human fall detection. One notable advantage of this approach is its suitability for implementation on edge devices, thanks to its low computational and memory requirements. This is achieved by integrating the previously untapped potential of the Fast Pose Estimation method, which had not been utilized for this specific purpose before. With its efficient utilization of resources and strong performance, the suggested technique holds significant promise in enhancing human fall detection systems.

The paper in [17] presented an approach to human fall detection in smart home environments by utilizing YOLO (You Only Look Once) networks. The research aims to enhance the safety and care provided in smart homes by addressing the critical issue of detecting human falls. The YOLO-based approach, specifically YOLOv5, is proposed as an efficient and accurate method for real-time fall detection. The paper discusses the development and implementation of this system, emphasizing its potential to improve the quality of care and response within smart home setups, thereby contributing to the broader field of healthcare and assisted living technology.

The paper in [18] developed a method for analyzing and deducing errors in human posture to mitigate musculoskeletal disorders among construction workers. The study employs

vision-based techniques to monitor and assess the postures of construction workers, aiming to identify and rectify potentially harmful positions. The findings suggest that this approach effectively reduces the risk of musculoskeletal disorders in the construction industry by providing real-time feedback and guidance on proper posture. However, a limitation of the study is not extensively discussing the practical implementation challenges and feasibility of the proposed vision-based system in real-world construction settings, which may require further investigation and adaptation.

The paper in [19] provided a comprehensive overview of the current state of research in vision-based indoor Human Activity Recognition. The review delves into the latest advancements in this field, highlighting the diverse applications and methodologies employed for recognizing human activities indoors using computer vision techniques. It discusses the challenges encountered, such as occlusions, variability in lighting conditions, and the need for large annotated datasets, which affect the accuracy and robustness of existing systems. The paper also presents future prospects, emphasizing the potential of deep learning models like Convolutional Neural Networks (CNNs) and recurrent networks to improve HAR accuracy, as well as the integration of multimodal sensor data to enhance performance. Overall, this review offers valuable insights into the current landscape of vision-based indoor HAR, pinpointing areas where further research and innovation are required to overcome existing challenges and unlock its full potential in various applications.

III. Methodology

YOLO was developed as a pretrained posture detector that can identify common items, including tables, chairs, automobiles, phones, and more. To develop a model capable of detecting human postures, such as those associated with walking, sitting, and falling. A detection technique based on YOLOv5 is what we suggest. Additionally, real-time is needed to detect and track the target [12]; therefore, our model performs well when used in real-time.

A. Dataset

For dataset preparation, images were collected from various sources, including Internet's webpages, and the Kaggle dataset [13], and created a custom posture dataset. This dataset involves images with three human posture classes: walking, sitting, and falling. The images directory contains two subdirectories (374 images) used for training and Val (111 images) for validation. The Labels directory contains two subdirectories, train and Val, and here in this directory, we have text files with labels for that particular image. Fig. 1 shows some examples of our dataset.

To enhance the performance of our model, we expanded the dataset from 374 images to 1092 images through augmentation using Roboflow. For each original image, up to five augmented versions were generated. These augmentations were applied randomly, involving rotations with hue variations ranging from -50° to +50°, adjustments in brightness within -40% to +40%, exposure changes between -35% to +35%, blurring up to 1 pixel, and the introduction of noise up to 5%-pixel value. Regarding the validation dataset, it consists of three distinct sets. The primary dataset remains unchanged, while the other two sets were obtained by applying different preprocessing techniques and augmentations.

In one of the preprocessing suites, the images were resized to 416x640, followed by automatic contrast adjustment using adaptive equalization, grayscale conversion, and incremental adjustments in brightness (between -40% and +40%) and exposure (between -21% and +21%). In another preprocessing set, the images were resized to 640x480, followed by automatic contrast adjustment using adaptive equalization. Additionally, brightness adjustments (between -40% and +40%), exposure alterations (between -21% and +21%), and 90° rotations (both clockwise and counterclockwise) were applied. For illustrative purposes, Fig. 2 showcases some examples of the augmented images resulting from these transformations.

Fig. 1. Sample images from the dataset.

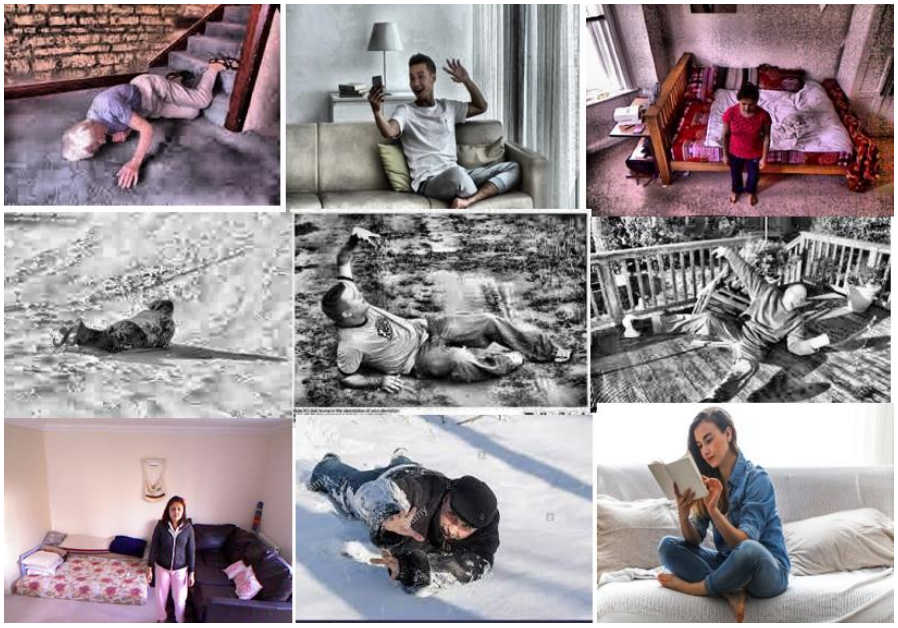

Fig. 2. Sample images of augmented images.

B. Google Colab

To conduct our experiments, we took advantage of the resourceful GPUs available through Google Colab, which offers free access to this powerful computing technology. Specifically, we utilized a 12GB NVIDIA Tesla T4 GPU for all our training and testing tasks. During the training phase, our model underwent 20 epochs, with a batch size 16, and the images were resized to a dimension of 640. Additionally, we maintained YOLOv5s's default settings for other hyperparameters to ensure consistency and fair comparison. For optimization purposes, we employed the Stochastic Gradient Descent (SGD) optimizer, setting the learning rate (lr) to 0.001, which further contributed to the effectiveness of our model during training.

C. Training and Testing

When embarking on the training of a posture detector, a highly effective strategy involves commencing with a pre-existing model that has been trained on extensive datasets. In this approach, the weights of this existing model are utilized as a starting point for training, even if the model's pre-trained weights do not directly encompass the specific postures required for the current experiment. This technique is commonly referred to as transfer learning. To expedite the learning process and enable faster convergence, we opted to utilize a pretrained model that incorporates weights previously trained on the COCO dataset. By leveraging this pretrained model as a starting point, our network can grasp and adapt to the novel posture detection task more swiftly and efficiently.

In this study, our total dataset consists of 1425 images, 70% of which are used for training, 20% for validation, and 10% for testing. 70% of training includes 1092 images, 20% of validation includes 333 images, and the rest of the images are considered for testing.

IV. Results and Analysis

In this section, we introduce the experiment's details, and then we show the training results using pretraining weights and compare the three models of YOLOv5. Then we validated the model on 333 photos. The experimental results adopt the average precision mean mAP and the number of frames per second (FPS). Correspondingly, for each category of postures, we calculated the Precision rate and Recall rate [2]. The results are shown in Fig. 3.

A. Performance Analysis

This section presents the performance analysis of Yolo models. The analysis justifies the usage of the Yolo algorithm and the specified version as YOLOv5. To perform this analysis, a comparison of Yolo models is presented.

When comparing different versions of YOLO models, accuracy and speed are two key metrics to consider. Accuracy refers to how well the model can detect and classify posture correctly. A more accurate model will have higher precision and recall in identifying postures in an image. Speed, on the other hand, relates to the inference time required for the model to process an image and provide the output. Faster models are desirable for real-time applications where quick posture detection is essential. Fig. 4 shows the comparing different versions of YOLO models on accuracy and speed [14].

As shown in Fig. 5, it's important to strike a balance between accuracy and speed, as a highly accurate model might sacrifice speed, while a faster model might compromise accuracy. Thus, when evaluating YOLO models, it is crucial to assess their performance based on accuracy and speed to choose the most suitable option for the intended use case. When comparing YOLO models based on accuracy and speed metrics, YOLOv5 stands out as a superior choice. YOLOv5 showcases significant improvements in both accuracy and speed compared to its predecessors, YOLOv3 and YOLOv4. Through optimized architecture and advanced training

strategies, YOLOv5 achieves better accuracy by accurately detecting and classifying postures in various scenes.

Additionally, YOLOv5 introduces model scaling options, allowing customization for specific requirements and striking a balance between accuracy and speed based on the task. Furthermore, its efficient inference techniques and streamlined design lead to faster processing times, making YOLOv5 ideal for real-time applications where quick posture detection is crucial. The overall enhancement in both accuracy and speed positions YOLOv5 as a top-performing posture detection model, making it a preferred choice for a wide range of computer vision applications.

Moreover, another graph is presented to represent the comparison of various YOLO models in terms of average precision (AP) while conducting the evaluation on YOLOv5l across different releases, YOLOv4 with different releases, and YOLOv3. The x-axis represents GPU time with a batch size of 8, indicating the time taken by the GPU to process the images. The y-axis represents the average precision (AP), which measures the accuracy of posture detection. Fig. 6 demonstrates the comparison of various YOLO models in terms of AP [15, 16].

As shown in Fig. 5, in analyzing the graph, it becomes evident that YOLOv5 consistently demonstrates better AP compared to YOLOv4 and YOLOv3. This indicates that YOLOv5 achieves more accurate posture detection across the evaluated releases. The higher AP scores attained by YOLOv5l across different releases suggest that its advancements in architecture, training strategies, and inference optimizations have significantly improved detection accuracy.

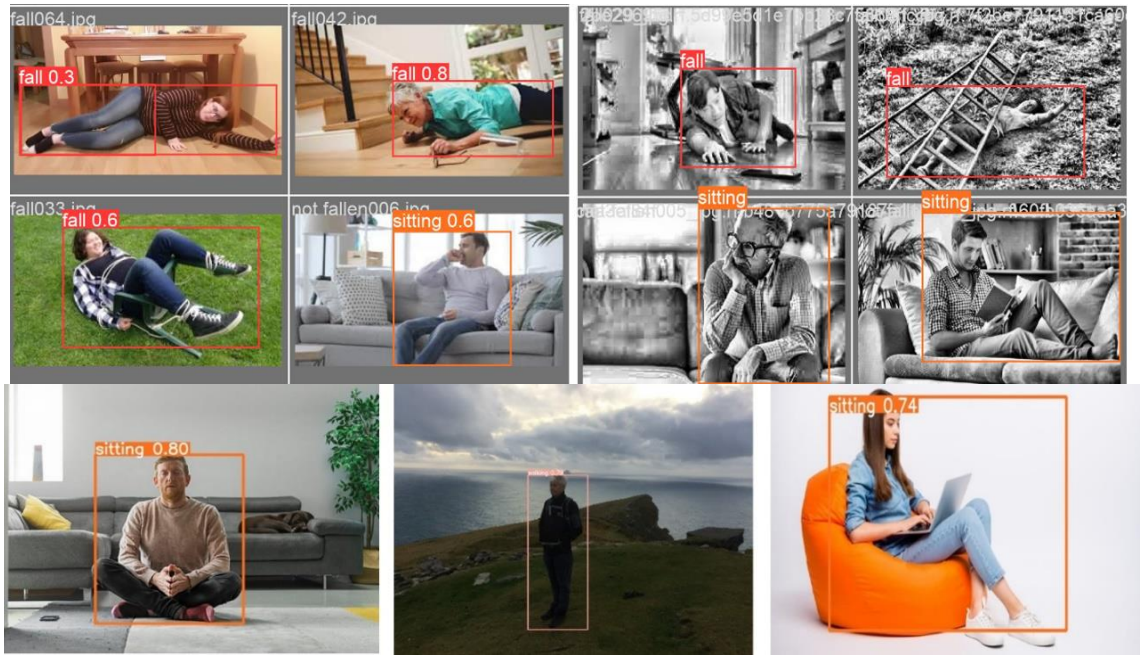

Fig. 3. Samples of experimental results.

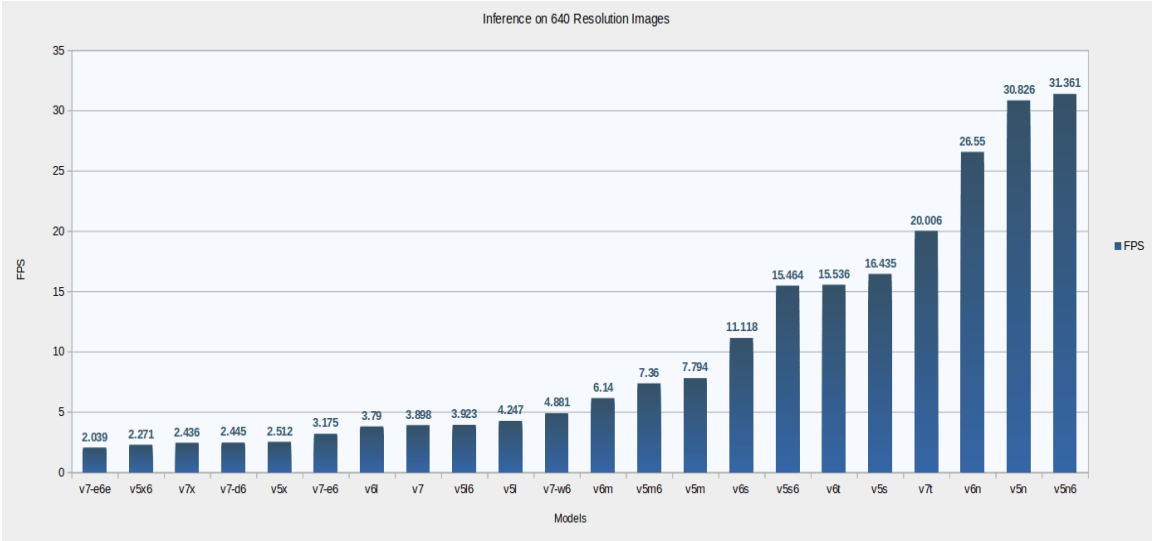

Fig. 4. Comparison of Yolo models based on accuracy and speed.

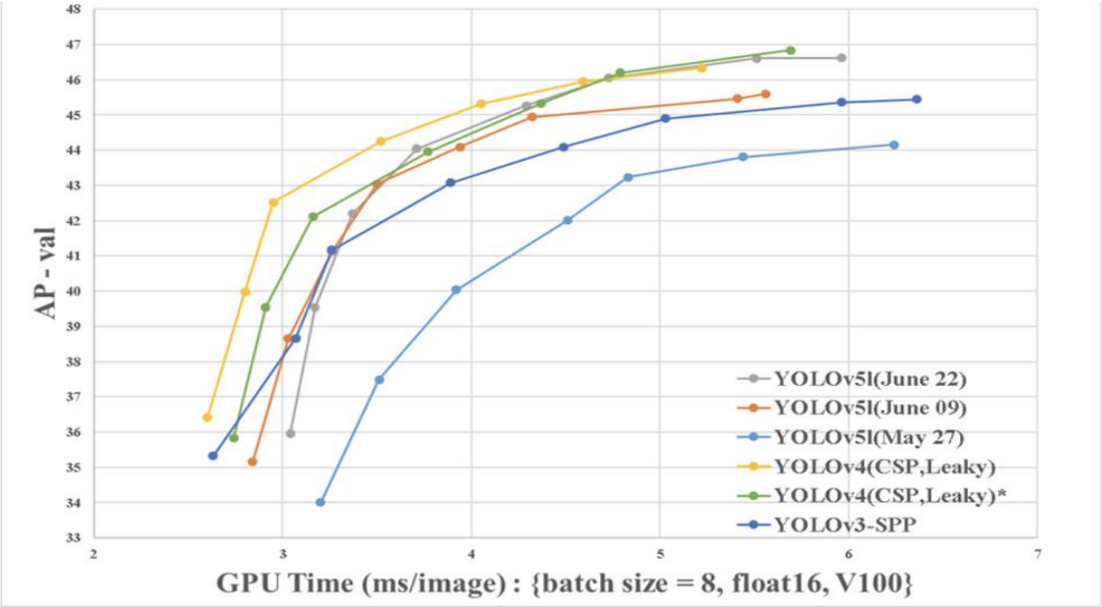

Fig. 5. Comparison of Yolo algorithm on average precision (AP).

Furthermore, the graph allows us to observe the trade-off between GPU time and AP for each model. By analyzing the trend lines or data points, it becomes apparent that YOLOv5 strikes a favorable balance between accuracy and GPU processing time. It manages to achieve higher AP scores while maintaining comparable or even faster GPU processing times compared to the other models.

Finally, the performance of the proposed method is evaluated based on precision, recall, and mAP. When evaluating the performance of a generated model for human posture detection, precision, recall, and mean Average Precision (mAP) are key metrics to consider. Precision measures the accuracy of positive predictions, indicating how well the model identifies true positives and minimizes false positives. Recall quantifies the model's ability to detect all relevant human postures, minimizing false negatives. These metrics help assess the model's accuracy and completeness in detecting human postures. Additionally, mAP combines precision and recall across various confidence thresholds, providing an overall measure of the model's performance. It allows for a comprehensive evaluation of the model's accuracy at different recall levels, offering insights into its performance across the entire range of detection thresholds.

By analyzing precision, recall, and mAP, one can gain a comprehensive understanding of the model's ability to accurately detect human postures while striking a balance between false positives and false negatives. Based on our experimental results, the first version of our model was trained with 1092 photos. This model achieved relatively good results. Table I shows the model evaluation in our proposed approach.

As shown in Table I, the table provides detailed performance metrics for a human posture detection model across different classes: "all," "fall," "walking," and "sitting." Let's examine each column as Class: This column represents the specific class or category of human postures for which the performance metrics are provided. Images: It indicates the number of images in the dataset that contain instances of the corresponding class. Instances: This column shows the total number of instances or occurrences of the specific class within the dataset. Precision: Precision measures the accuracy of the model's positive predictions for each class. It indicates the proportion of correctly identified instances out of all the predicted instances for that class. Higher precision values indicate a lower rate of false positives. Recall: Recall, also known as sensitivity, represents the model's ability to detect all relevant instances of the class. It calculates the proportion of correctly identified instances out of all the actual instances for that class. Higher recall values indicate a lower rate of false negatives. mAP50: (mAP at IoU threshold 0.50) evaluates the overall detection accuracy of the model. It measures the average precision across all classes at a specific Intersection over the Union (IoU) threshold of 0.50. Higher mAP50 values indicate better overall detection performance. mAP50-90: mAP50-90 represents the mean Average Precision averaged across all classes, but with IoU thresholds ranging from 0.50 to 0.90. This metric provides a broader assessment of the model's performance by considering a range of IoU thresholds.

As depicted in Fig. 6, analyzing the table, we can observe the overall performance of the model, represented by the "all" class, achieves a precision of 0.814, recall of 0.794, mAP50 of 0.825, and mAP50-90 of 0.529. Among the specific classes, the "fall" class achieves the highest precision of 0.909, recall of 0.834, mAP50 of 0.924, and mAP50-90 of 0.539. This indicates that the model performs particularly well in detecting instances of falling postures. The "walking" class shows a high recall of 0.899, indicating that the model effectively detects walking postures while achieving a precision of 0.827 and mAP50 of 0.886. The "sitting" class exhibits lower precision (0.705) and recall (0.649), suggesting that the model faces challenges in accurately detecting instances of sitting postures. It achieves a mAP50 of 0.664, indicating moderate overall performance.

TABLE I. MODEL EVALUATION METRICS

Class	Images	Instances	Precision	Recall	mAp50	Map50-90
all	333	342	0.814	0.794	0.825	0.529
fall	333	216	0.909	0.834	0.924	0.539
walking	333	69	0.827	0.899	0.886	0.619
sitting	333	57	0.705	0.649	0.664	0.429

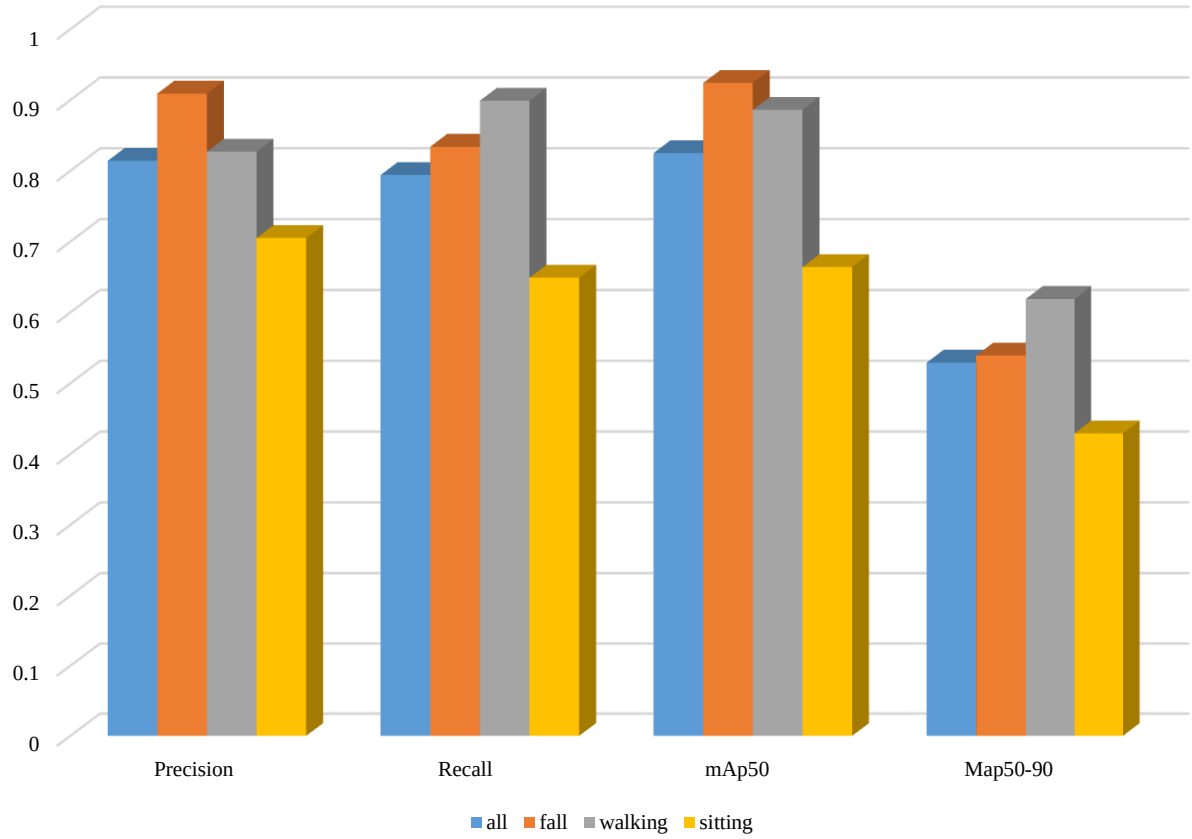

Fig. 6. Performance analysis of our generated model.

V. CONCLUSION

Human posture detection is a challenging task in smart home applications. This study deals with the high complexity, low accuracy, and speed challenges of human posture detection in smart home applications. It intends to develop a vision-based posture recognition system to utilize in smart home applications that can identify human posture. In the proposed approach, a YOLO network is trained and tested to recognize the postures in our custom dataset. This Yolo-based approach is based on YOLOv5, which provides a high accuracy rate and satisfied speed in posture detection. Experimental results show the effectiveness of the developed system for posture recognition on smart home applications. Future research directions could involve refining real-time detection capabilities and exploring user-centric applications, expanding the utility of posture recognition beyond health and safety monitoring to enhance user comfort and experience in smart home environments.

ACKNOWLEDGMENT

This word was supported by: Research on Smart Home System Based on Big Data Environments (JF-LX202001)

REFERENCES

[1] N. Yu, J. Lv, Human body posture recognition algorithm for still images, The Journal of Engineering, 2020 (2020) 322-325.

[2] W. Quan, J. Woo, Y. Toda, N. Kubota, Human posture recognition for estimation of human body condition, Journal of Advanced Computational Intelligence and Intelligent Informatics, 23 (2019) 519-527.

[3] A. Aghamohammadi, M.C. Ang, E. A. Sundararajan, N.K. Weng, M. Mogharrebi, S.Y. Banihashem, A parallel spatiotemporal saliency and discriminative online learning method for visual target tracking in aerial videos, Plos one, 13 (2018) e0192246.

[4] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, D.D. Feng, Deep convolutional neural networks for human action recognition using depth maps and postures, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49 (2018) 1806-1819.

[5] G. Liu, L. Lin, W. Zhou, R. Zhang, H. Yin, J. Chen, H. Guo, A Posture Recognition Method Applied to Smart Product Service, Procedia CIRP, 83 (2019) 425-428.

[6] M.C. ANG, A. AGHAMOHAMMADI, K.W. NG, E. SUNDARARAJAN, M. MOGHARREBI, T.L. LIM, MULTI-CORE FRAMEWORKS INVESTIGATION ON A REAL-TIME OBJECT TRACKING APPLICATION, Journal of Theoretical & Applied Information Technology, 70 (2014).

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

[8] S. Sivamani, S.H. Choi, D.H. Lee, J. Park, S. Chon, Automatic posture detection of pigs on real-time using Yolo framework, Int. J. Res. Trends Innov, 5 (2020) 81-88.

[9] M. Dallel, V. Havard, Y. Dupuis, D. Baudry, A Sliding Window Based Approach With Majority Voting for Online Human Action Recognition using Spatial Temporal Graph Convolutional Neural Networks, 2022 7th International Conference on Machine Learning Technologies (ICMLT), 2022, pp. 155-163.

[10] Z. Zhang, X. Ma, H. Wu, Y. Li, Fall detection in videos with trajectory-weighted deep-convolutional rank-pooling descriptor, IEEE Access, 7 (2018) 4135-4144.

[11] M. Salimi, J.J. Machado, J.M.R. Tavares, Using deep neural networks for human fall detection based on pose estimation, Sensors, 22 (2022) 4544.

[12] M. Ang, E. Sundararajan, K. Ng, A. Aghamohammadi, T. Lim, Investigation of Threading Building Blocks Framework on Real Time Visual Object Tracking Algorithm, Applied Mechanics and Materials, 666 (2014) 240-244.

[13] U.K. KANDAGATLA, Fall Detection Dataset, 2021.

[14] Sovit Rath, V. Gupta, Performance Comparison of YOLO Object Detection Models – An Intensive Study, 2022.

[15] J. Solawetz, What is YOLOv5? A Guide for Beginners., 2020.

[16] G. Jocher, Ultralytics - yolov5, 2023.

[17] Bo LU. Human Fall Detection for Smart Home Caring using Yolo Networks. International Journal of Advanced Computer Science and Applications. 2023;14(4).

[18] Purushothaman MB, Gedara KM. Smart vision-based analysis and error deduction of human pose to reduce musculoskeletal disorders in construction. Smart and Sustainable Built Environment. 2023.

[19] Bhola G, Vishwakarma DK. A review of vision-based indoor HAR: state-of-the-art, challenges, and future prospects. Multimedia Tools and Applications. 2023 May 11:1-41.