A practical and robust skeleton-based artificial intelligence algorithm for multi-person fall detection on construction sites considering occlusions

Doil Kim[a,b], Xiaoqun Yu[c], Shuping Xiong[a,*]

[a] Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea.

[b] AI Engineering Lab, LG CNS, Seoul 07795, Republic of Korea

[c] Department of Industrial Design, School of Mechanical Engineering, Southeast University, Nanjing 211189, China.

*Corresponding author: Prof. Shuping Xiong; Tel: +82-42-350-3132; Email: shupingx@kaist.ac.kr

E-mail addresses: gjgjos@naver.com (D. Kim), xiaoqunyu@seu.edu.cn (X. Yu), shupingx@kaist.ac.kr (S. Xiong).

Abstract

Rapid and accurate fall detection is crucial for enhancing worker safety and mitigating the severity of fall-related incidents on construction sites. This research tackles the challenge of real-time, accurate multi-person fall detection in these dynamic and obstructed environments. To achieve this, we constructed a comprehensive fall video dataset that encompasses diverse fall scenarios specific to construction sites. We also proposed a novel, robust skeleton-based artificial intelligence algorithm for multi-person fall detection, which leverages the strengths of computer-vision-based human detection, tracking, and extraction and analysis of multi-person skeleton keypoints across consecutive frames. The proposed model outperforms existing state-of-the-art fall detection algorithms, achieving an impressive accuracy of 98.66% (sensitivity: 97.32%; specificity: 99.10%). Importantly, our model demonstrates robust performance under varying occlusion levels, showcasing the efficacy of the proposed algorithm in maintaining detection accuracy irrespective of occluded conditions. In addition, we have deployed the algorithm on an edge device, where the system ran at 6.44 frames per second, meeting the requirements for real-time performance in portable monitoring applications. With its high robustness and practicality, the proposed artificial intelligence algorithm shows significant potential for real-world fall detection, thereby enhancing worker safety in the construction industry.

Keywords: fall detection, construction safety, pose estimation, deep learning, edge computing

1 Introduction

Falls are the leading cause of injuries and deaths at construction sites (Choi et al., 2019; Hu et al., 2011). According to the statistics, falls are the most common cause of occupational accidents in the construction industry, accounting for 33% of all accidents in Korea (South Korea Ministry of Employment and Labor, 2020), and over 47% of all fatal construction industry incidents in Hong Kong (Chan et al., 2008). In addition to physical injuries, falls from heights can also cause mental trauma (Agarwal et al., 2020). Therefore, prompt detection and response to fall incidents are essential for mitigating injury severity and preventing fatalities. Several studies highlight the significance of fall detection systems in identifying falls and initiating appropriate actions (Fanca et al., 2019; Newaz & Hanada, 2023; Schröter et al., 2023). These systems could enable immediate medical attention, which can be crucial during the golden hour after an accident, potentially saving lives and reducing long-term disability.

Traditional fall detection systems often rely on dedicated safety supervisors who monitor live video feeds for potential fall incidents. This approach is not only costly but also prone to human error (Pan & Zhang, 2021). With advancements in artificial intelligence (AI), AI-based safety technologies have been proposed to address the limitations of conventional manual observation methods in construction monitoring and management (Pan & Zhang, 2021). Automated monitoring techniques for fall detection can be divided into two main categories: sensor-based methods and computer vision (CV)-based methods (Nath et al., 2020; Sarkar et al., 2020). Kim et al. (Y. Kim et al., 2020) utilized an inertial measurement unit (IMU) sensor for detecting pre-impact falls from heights based on a threshold-based algorithm. Abbate et.al (Abbate et al., 2012) developed a smartphone-based fall detection system that leverages an accelerometer to identify falls by movement peaks and automatically sends help requests. While the sensor-based approach offers stable performance across various weather conditions, its implementation requires a substantial long-term investment in purchasing, installing, and maintaining complex sensor networks (Nath et al., 2020; Wang et al., 2021). Additionally, the requirement for individuals to wear sensors can lead to discomfort and thus hinder compliance (Z. Zhang et al., 2015).

On the contrary, CV-based techniques have been discovered to be cost-effective for remotely monitoring construction projects without invasive methods (Pal & Hsieh, 2020). Additionally, CV-based methods hold the advantage of providing comprehensive information such as workers' locations, behaviors, and site conditions. This facilitates a deeper understanding of complex construction tasks than sensor-based methods, which offer limited information (Seo et al., 2015). Numerous studies have explored CV-based deep learning models for fall detection. Unlike traditional methods which rely on hand-crafted features, deep learning has become the mainstream approach due to its powerful ability to automatically extract features. Yu et.al (M. Yu et al., 2017) adopted a convolutional neural network (CNN)-based method for fall detection in elderly care, utilizing silhouette extraction from video streams. Similarly, Leite et.al (Leite et al., 2019) proposed VGG-16 with support vector machine (SVM) classifiers, integrating optical flow, saliency maps, and RGB data for fall detection among the elderly. Zheng et.al (Y. Zheng et al., 2019) applied a graph convolution network (GCN) to keypoints extracted by OpenPose (Cao et al., 2017), effectively minimizing the influence of appearance factors. Recently, skeleton-based action recognition has emerged as a prominent approach, particularly for its robustness in dynamic, complex, and occluded

environments (Du et al., 2015; Bai et al., 2023). This technique, which represents human movement through 3D coordinates of body joints, has received significant focus for its ability to create efficient and stable models that are less affected by variables like lighting and body dimensions, making it a promising method (Rouali et al., 2023).

Fall detection at construction sites faces unique challenges due to the inherently dynamic and obstructed environments, coupled with the presence of multiple workers engaged in high-risk activities often at heights. These specific conditions remain insufficiently addressed by the majority of current research, which predominantly focuses on fall detection in controlled or eldercare-specific settings. As shown in **Table 1**, there are various video datasets related to falls such as multiple camera dataset (Auvinet et al., 2010), UP-fall dataset (Martínez-Villaseñor et al., 2019), UR-fall dataset (Kwolek & Kepski, 2014), Le2i fall dataset (Charfi et al., 2013), SisFall dataset (Sucerquia et al., 2017). These datasets do not consider falls from heights – a critical aspect of construction safety, and commonly feature single-person scenarios in controlled environments that bear little resemblance to the chaotic and complex conditions of construction sites.

Existing efforts to refine fall detection algorithms, such as Baldewijns et.al (Baldewijns et al., 2016) and An et.al (An et al., 2021) typically simulate environments such as nursing homes or cover general real-world scenarios that do not adequately encapsulate the challenges present at construction sites. Zhang et.al (Y. Zhang et al., 2023) introduced a fall detection model that combines an improved Spatial-Temporal Graph Convolutional Network (ST-GCN) with the BlazePose algorithm (Bazarevsky et al., 2020), but their model falls short in multi-person tracking. Similarly, Inturi et.al (Inturi et al., 2023) proposed a vision-based fall detection model employing AlphaPose and long short-term memory (LSTM), but it cannot also track multiple individuals. While Feng et.al (Feng et al., 2020) developed an attention-guided LSTM model that effectively locates and tracks multiple individuals using a combination of YOLOv3 (Wojke et al., 2017), Deep-Sort (Redmon & Farhadi, 2018), and CNN for crowded scenes, its real-world applicability outside laboratory settings considering occlusion remains untested.

Table 1. CV-based public datasets for fall detection.

Dataset	Detection type	Falls from heights	Action type	Videos
NTU RGB+D 120 (Liu et al., 2019)	Multi-person (2)	No	120 (Daily actions, falling)	0.1M videos (total) 948 videos (only for falling)
Custom dataset (Baldewijns et al., 2016)	Single-person	No	2 (ADL, Fall)	55 scenarios
Multiple camera dataset (Auvinet et al., 2010)	Multi-person (>2)	No	2 (ADL, Fall)	24 scenarios
Up-fall dataset (Martínez-Villaseñor et al., 2019)	Single-person	No	11 (ADL 6, Fall 5)	17 subjects with 11 activities
UR fall dataset (Kwolek & Kepski, 2014)	Single-person	No	2 (ADL, Fall)	70 videos

Le2i fall dataset (Charfi et al., 2013)	Single-person	No	2 (ADL, Fall)	191 videos
SisFall dataset (Sucerquia et al., 2017)	Single-person	No	2 (ADL, Fall)	34 videos
CAUCAFall dataset (Guerrero et al., 2022)	Single-person	No	2 (ADL, Fall)	10 subjects with 10 activities
Fall detection dataset (Adhikari et al., 2017)	Single-person	No	6 (Standing, sitting, Lying, bending, crawling, empty)	20 videos
VFP290K (An et al., 2021)	Multi-person (>2)	No	2 (Fallen, Non-fallen)	178 videos
Custom dataset (Gomes et al., 2022)	Multi-person (>2)	No	2 (Fall, Non-fall)	Include 430 movie clips

Recognizing these critical gaps, our study introduces a sophisticated yet practical approach tailored specifically to the complexities of construction site monitoring. By harnessing the power of edge computing, our solution significantly reduces dependency on cloud computing, which is crucial in locations with fluctuating connectivity and limited network access. This localized processing not only enhances data privacy but also conserves bandwidth, essential for maintaining uninterrupted service across expansive operational fields (Gallo et al., 2022; Nguyen et al., 2021).

Our approach utilizes a multi-person fall detection system that operates effectively in real-time on edge devices. This system minimizes reliance on cloud computing, ensuring continuous operation even in disconnected or network-limited scenarios. Such a feature is particularly vital for ensuring worker safety in parts of construction sites where traditional monitoring systems fail due to occlusions or the extensive areas they need to cover. Moutsis et.al (Moutsis et al., 2023) and Zheng et.al (H. Zheng & Liu, 2022) proposed lightweight fall detection systems on portable devices like the Raspberry Pi and Jetson Nano, achieving notable frame rates and accuracy. However, these models were not tested under conditions that simulate the challenging and variable environments of construction sites. Change et.al (Chang et al., 2021) introduced a pose estimation-based fall detection system using OpenPose-light, centroid tracker, and LSTM on Jetson TX2, demonstrating portable deployment potential. Nevertheless, their model benchmarked against the indoor-focused Le2i dataset (Charfi et al., 2013), did not account for the dynamic features of outdoor construction sites.

Our study addresses the scarcity of research on real-time, skeleton-based multi-person fall detection using edge devices in dynamic environments of construction sites. We aim to develop a novel, real-time multi-person fall detection model called YOSAP-LSTM on edge devices, integrating advanced deep learning models based on a large-scale fall dataset. This model combines YOLOv8 for person detection, AlphaPose and SORT algorithms for multi-person skeleton keypoint extraction across frames, and 1D CNN-LSTM for classifying falls and non-falls. Unique to our approach is the validation of the fall detection model according to different levels of occlusion, a first in this field. The main contributions of this study are as follows:

1) We construct a comprehensive large-scale fall video dataset capturing diverse scenarios, essential for developing accurate fall detection systems;

2) We propose a novel skeleton-based multi-person fall detection model that integrates state-of-the-art deep learning techniques for accurate real-time fall detection on edge devices;

3) We develop a new portable AI-based safety monitoring system to detect fall incidents on edge devices with high and robust accuracy adaptable to different occlusion levels and meeting real-time FPS for remote flexible monitoring at construction sites.

The methods used to create a fall video dataset and integrate the YOLOv8, SORT, AlphaPose, and 1D CNN-LSTM for skeleton-based multi-person fall detection, along with the experimental setup, evaluation metrics, and main results, are reported in the following sections. The system is designed to be flexible and adaptable for remote monitoring of the dynamic and complex conditions of construction sites.

2 Methods

2.1 Dataset construction

Constructing a comprehensive and high-quality dataset is important for training deep learning models, as it directly affects model performance and generalization to real-world problems. For training YOLOv8 to detect individuals, the raw images are collected from the public dataset, field dataset, and web as shown in **Figure** 1. The public datasets utilized in this study include the Roboflow hard hat workers dataset (Roboflow, 2022) and the Ai-Hub construction site safety equipment dataset (AI-HUB Dataset, 2020) as both contain abundant images of construction site workers, providing realistic scenarios of laborers in action. The Roboflow dataset and Ai-Hub dataset only have annotations for personal protective equipment, requiring additional annotations for person class. In terms of constructing the dataset, various angles, shapes, and brightness were considered ensuring the model's ability to generalize across different construction environments. The field dataset was acquired using a mobile phone camera to capture images from construction sites and web-based images are crawled from Google.

(a) (b) (c)

Figure 1. Examples of raw images from different sources. (a) public dataset, (b) field dataset, (c) web crawling.

The image dataset was preprocessed and labeled manually. All images were resized to 640x640 resolution and labeled into a single class using the YOLO-Label annotation tool

(Yonghye Kwon, 2019). **Figure 2** shows examples of annotated images. With a significant count of 24,264 'person' instances in the total dataset, the deployment of the YOLOv8 allowed for the precise extraction of worker bounding boxes. The dataset is divided into a training set (70%), a validation set (15%), and a test set (15%), resulting in 17,950 instances for training, 3,377 instances for validation, and 2,937 instances for testing.

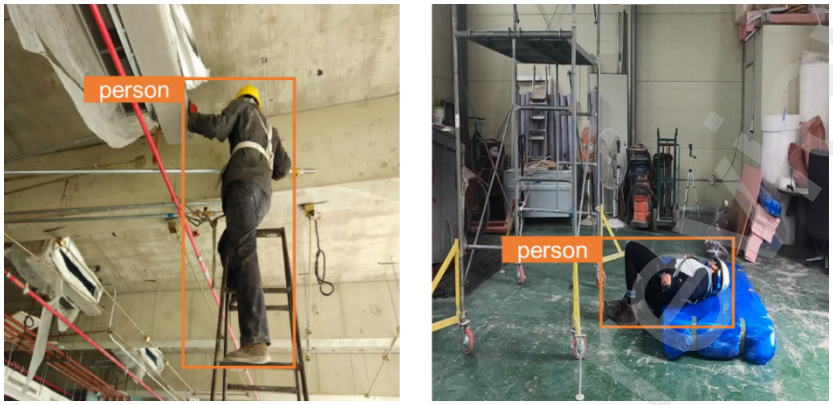
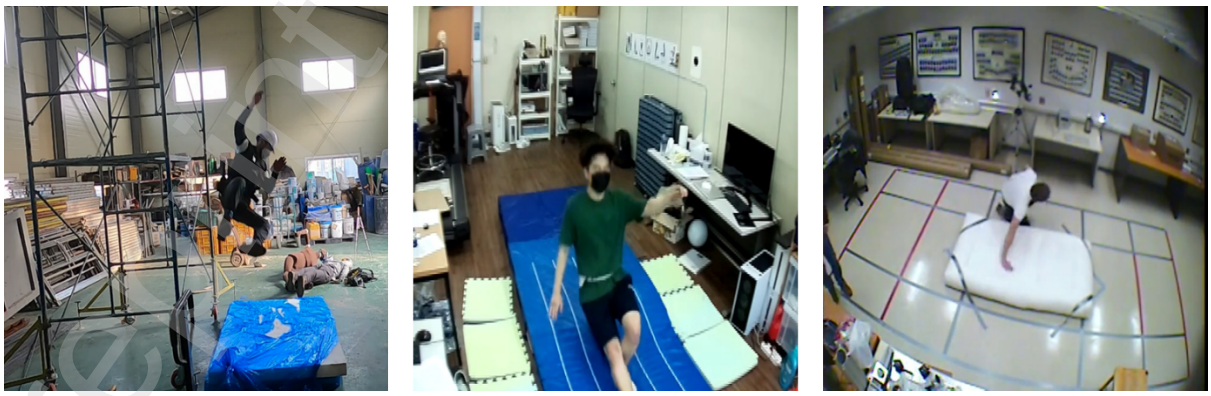


Figure 2. Examples of annotated images.

To train the fall classification model, we constructed a comprehensive fall video dataset comprising 376,006 frames, derived from three distinct sources, as shown in **Figure 3**. The KFall dataset (X. Yu et al., 2021), generated in indoor environments with 323,412 frames, and the multiple camera dataset (Auvinet et al., 2010), with 14,678 frames captured from various viewpoints, offer a diverse range of fall incidents. The Ai-Hub dataset (AI-HUB Dataset, 2020), sourced from actual construction site simulations, contributes 35,385 frames, providing realistic fall scenarios. Each dataset originally varied in resolution – with the KFall dataset at 640x480, the multiple camera dataset at 720x480, and the Ai-Hub dataset at 1920x1080. For consistency, the resolution across all videos is standardized to 640x480 pixels. Each frame is labeled as 'Non-fall' (0) or 'Fall' (1). The definition of a 'Fall' includes both the beginning of the falling motion and the state post-fall, ensuring comprehensive coverage of the fall event spectrum in our analysis. To train the fall classification model, the dataset is divided into a training set (80%) and a validation set (20%).



Ai-Hub dataset KFall dataset Multiple camera dataset

Figure 3. Examples of fall datasets.

2.2 Overall multi-person fall detection model

We develop a skeleton-based real-time multi-person fall detection model, YOSAP-LSTM, designed for monitoring multiple workers on a construction site, using the edge device as shown in **Figure 4**. The model operates by initially acquiring real-time video feeds from placed edge devices and cameras, which are then processed using the YOLOv8. The model employs a SORT (Simple Online and Realtime Tracking) algorithm (Bewley et al., 2016) to maintain consistent tracking of each worker across the video frames. Following this, the AlphaPose (H.-S. Fang et al., 2022) is applied to each identified worker to extract precise skeleton keypoint data. These keypoints, captured over successive frames, are fed into a combined 1D-CNN (one-dimensional convolutional neural network) and LSTM (long short-term memory) model, a robust time series analysis architecture, to accurately classify whether the observed movements are a fall or non-fall.

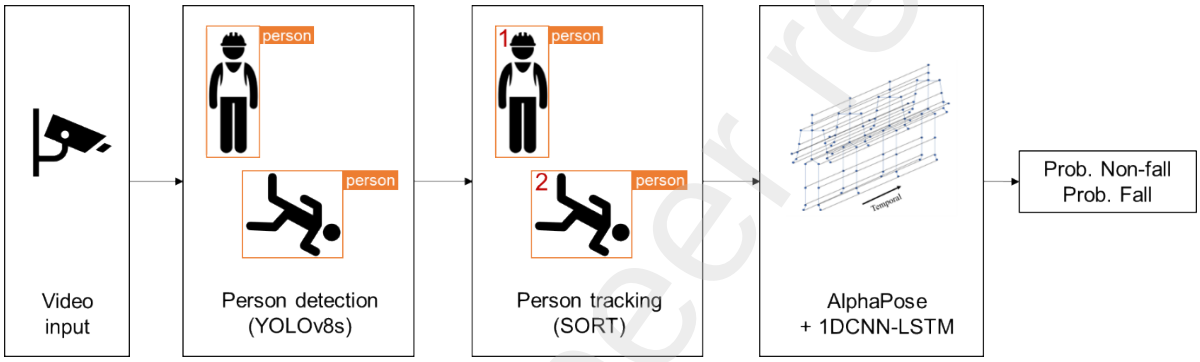


Figure 4. The schematic diagram of our proposed YOSAP-LSTM model.

2.2.1 Person detection and tracking

For detecting a person, we leverage the advanced capabilities of the YOLOv8 to detect workers. YOLO (You Only Look Once) (Redmon et al., 2016) is a real-time object detection algorithm developed by Redmon et al. in 2016. As a one-stage object detector, YOLO has fast inference speed as it predicts bounding boxes and object class probability simultaneously. YOLOv8 (Jocher et al., 2023), the latest YOLO model developed by Jocher et al. in 2023, is selected in this study. The YOLOv8s, the lightweight version of YOLOv8, is chosen as the baseline model because it is suitable for deployment on the edge device. We fine-tuned YOLOv8 using the custom dataset. After training the YOLOv8, it becomes adept at recognizing human figures on a construction site. With a significant count of 24,264 'person' instances in the total dataset, the deployment of the YOLOv8 allowed for the precise extraction of worker bounding boxes.

To track multiple workers in construction sites, SORT(Bewley et al., 2016), an algorithm used for object tracking in video frames, is employed. It is designed to be efficient, requiring minimal computational resources, and operates in real-time. The algorithm is particularly useful in scenarios that require fast and accurate object tracking, such as surveillance, robotics, and autonomous vehicles. In this system, YOLOv8 detects workers and generates bounding boxes, which SORT then uses to maintain consistent object trajectories across frames as shown in **Figure 5**. This combination effectively handles the dynamics of construction site environments, managing occlusions and the varied movements of workers with minimal computational overhead.

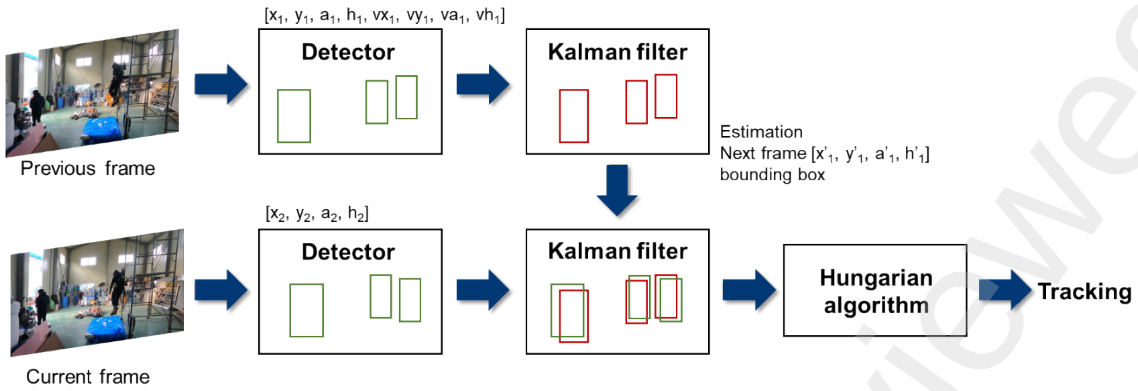


Figure 5. The schematic diagram of the SORT algorithm.

2.2.2 Skeleton-based fall classification

We incorporate AlphaPose (H.-S. Fang et al., 2022), a real-time multi-person pose estimation system, to enhance our fall detection model. It is known for its accuracy and efficiency, making it suitable for a variety of applications such as human-computer interaction, augmented reality, sports analysis, and video surveillance. AlphaPose is utilized to accurately detect the skeletal structure of workers on construction sites, which is crucial for assessing postures that may indicate a fall. By integrating this technology with our YOLOv8-based detection system, we can pinpoint and analyze the positions of key body joints, enabling precise recognition of fall-related movements and poses. This combination ensures robust and efficient monitoring of workers, improving safety management by providing timely alerts in case of fall incidents.

To enhance efficiency and minimize computational costs, we modify the default AlphaPose configuration by retaining only the nose from the facial keypoints – excluding the eyes and ears – and thus use 13 joints for pose estimation. This selective approach retains all critical joints for a full pose representation while optimizing processing time. **Figure 6** presents these chosen joints - Nose, Left Shoulder, Right Shoulder, Left Elbow, Right Elbow, Left Wrist, Right Wrist, Left Hip, Right Hip, Left Knee, Right Knee, Left Ankle, and Right ankle. The resulting keypoints are extracted as (x,y) coordinates with associated confidence scores.

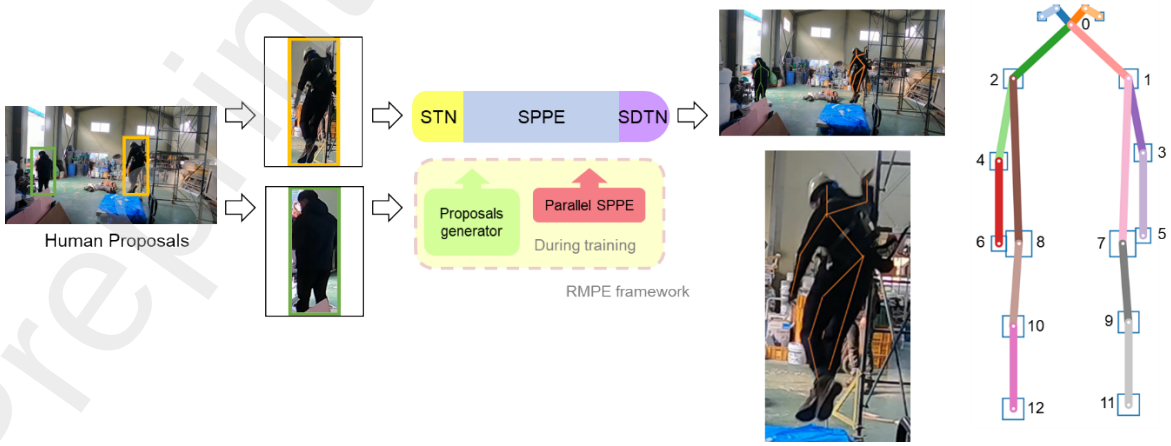


Figure 6. The schematic diagram of the AlphaPose and annotated 13 skeleton keypoints.

Following the acquisition of skeleton keypoints, which include each body joint across sequential video frames as illustrated in **Figure 7**, the 1D CNN-LSTM hybrid model is employed to effectively process a sequence of keypoints over time in the context of action recognition. A 1D CNN-LSTM hybrid model combines the strengths of both Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) to analyze sequence data, utilizing CNN's ability to extract spatial features and LSTM's proficiency in capturing temporal dependencies.

To process the sequences, we define a window size of 30 frames, equating to one second of video, and employ a sliding window technique with a step size of 10 frames to handle consecutive frames as a unified batch. The majority label within one window determines its assigned label for a window. Addressing the instances where AlphaPose fails to detect a human for some frame, we use cubic spline interpolation (McKinley & Levine, 1998) for up to ten missing frames, creating a continuous estimate of movement trajectories for augmenting the training dataset. Cubic interpolation is a type of spline interpolation that fits a cubic polynomial to the data points in such a way that the resulting curve appears smooth and continuous. Cubic interpolation can be beneficial when dealing with time series data that exhibits non-linear patterns or when the goal is to estimate the underlying function that could have generated the observed data points with minimal error (Castellano, 2023).

The 1D CNN layer is responsible for extracting local patterns and features from the input sequence (Chandrasekaran & Kumar Paramasivan, 2022a). Since the input is a sequence of joint keypoints, a 1D convolution is applied along the temporal dimension to capture the spatial relationships between joints across frames. The CNN layer applies convolutional filters to the input sequence, capturing local dependencies and extracting features that are crucial for understanding the human pose and activity. The convolutional layer can also help in reducing the dimensionality of the input data, making it more manageable for subsequent processing. Following the 1D CNN layer, the extracted features are fed into an LSTM layer. LSTM (Staudemeyer & Morris, 2019) is a type of recurrent neural network (RNN) that can capture long-term dependencies in sequential data, providing a comprehensive understanding of the temporal dynamics of the human pose. It models the sequential nature of the data, providing a context for each frame in relation to the previous frames. Upon extracting features through the CNN and LSTM layers, a fully connected layer outputs the probability of fall and non-fall for binary classification. The hybrid model is robust to variations in pose and can handle sequences of varying lengths due to the windowing approach. To fine-tune our model, we employ focal loss (T. Y. Lin et al., 2017), which is particularly effective when training with imbalanced datasets – like our dataset, where ‘non-fall’ samples outnumber ‘fall’ samples. It allows the model to concentrate learning on difficult cases that are often underrepresented in the training data, potentially improving the model's sensitivity to actual fall events.

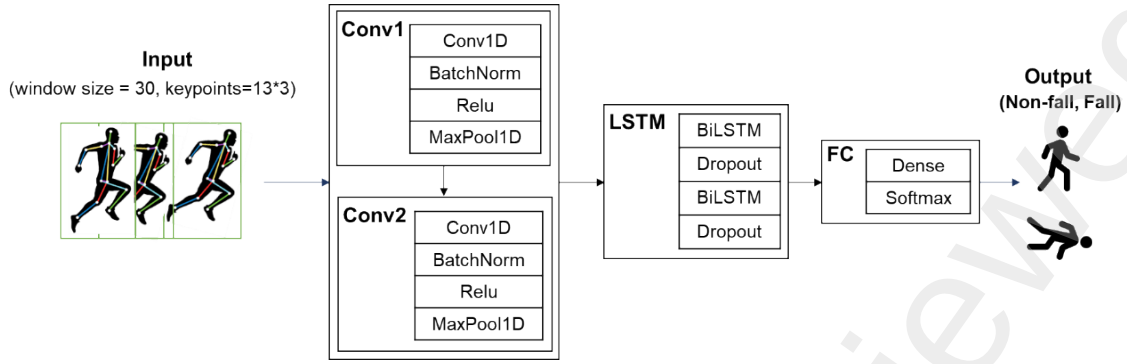


Figure 7. The schematic diagram of 1D CNN-LSTM.

2.3 Environmental setup

The training process was conducted on a local computer with a 12-core Intel® Core (TM) i7-12700K CPU clocked at 3.61 GHz and 96 GB of memory, accompanied by a NVIDIA GeForce RTX 3070 Ti GPU with 8 GB video memory. The operating system was Windows 10, and the deep learning model framework used PyTorch 1.11.0, with Python version 3.8.16 and CUDA version 11.3. The training parameters of 1D CNN-LSTM are shown in **Table 2**. The trained models were deployed on the NVIDIA Jetson Xavier NX for calculating FPS. The NVIDIA Jetson Xavier NX has an embedded NVIDIA Volta GPU core with 48 Tensor Cores, 384 CUDA cores, and 8 GB of memory.

Table 2. Training parameters of 1D CNN-LSTM.

Parameter	Value	Parameter	Value
Epochs	100	Window size	30
Learning rate	0.0001	Input channels	39
Batch size	64	Optimizer	Adam

2.4 Evaluation metrics

The performance of the fall classification model is evaluated by three primary metrics: accuracy, sensitivity, and specificity. The equations about accuracy, sensitivity, and specificity are calculated by equation (1), (2), and (3), respectively,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

where TP is true positive, FP is false positive, FN is false negative, and TN is true negative.

Accuracy is the most intuitive performance measure, and it gives a baseline comparison. Sensitivity, also known as recall or true positive rate, measures the proportion of actual falls that are correctly identified by the model. It is crucial for fall detection systems as it reflects the model's ability to detect potentially dangerous events. A model with high sensitivity is less likely to miss fall events, thereby reducing the risk of not triggering an alert when a fall occurs. Specificity, on the other hand, indicates the proportion of non-fall events that are correctly identified by the model. High specificity means that the model is good at avoiding false alarms, which can be particularly important in maintaining user trust and reducing the operational overhead of dealing with false alerts.

Additionally, the model's processing speed is evaluated by Frames Per Second (FPS) to guarantee its applicability in real-time situations. The ability to rapidly and accurately detect and track worker movements and activities is vital in identifying risks such as fall incidents, ensuring timely interventions in critical situations on construction sites. In this context, FPS is a key performance metric, with higher FPS ensuring smoother video playback, facilitating prompt detection and response to hazards. While high-speed tasks like traffic monitoring require over 10 FPS, less dynamic tasks may need only 2-3 FPS (Lee & Hwang, 2022). Our study, in line with previous studies (Nath et al., 2020; Ren et al., 2017), considers above 5 FPS as adequate for real-time performance in fall detection.

3 Results

3.1 Model performance and comparison with previous models

Before evaluating the performance of the 1D CNN-LSTM model, the initial step in our process involves the application of AlphaPose for person detection. The initial person detection accuracy is crucial, as it sets the baseline for the subsequent fall classification. The accuracy of person detection varies across datasets, with the total dataset achieving 93.60%, the KFall dataset at 95.94%, the multiple camera dataset at 86.19%, and the Ai-Hub dataset at 76.21%. The relatively lower accuracy in the Ai-Hub dataset (76.21%) implies that this dataset might contain more challenging scenarios for person detection compared to the KFall dataset (95.94%). This could be due to factors like rapid fall motions, greater occlusion, diverse activities, and varied postures and orientations.

After ensuring accurate detection, we conducted a comparative analysis of the 1D CNN-LSTM model against various other models. Following Chen et.al (Chen et al., 2020), as shown in **Figure 8**, we evaluated a threshold-based model involving two criteria: firstly, the angle between the body's centerline (derived from the arctan calculation between the nose and the midpoints of the right and left hips) and the ground must be less than 60˚; secondly, the width-to-height ratio of the body's bounding box should exceed 2/3.

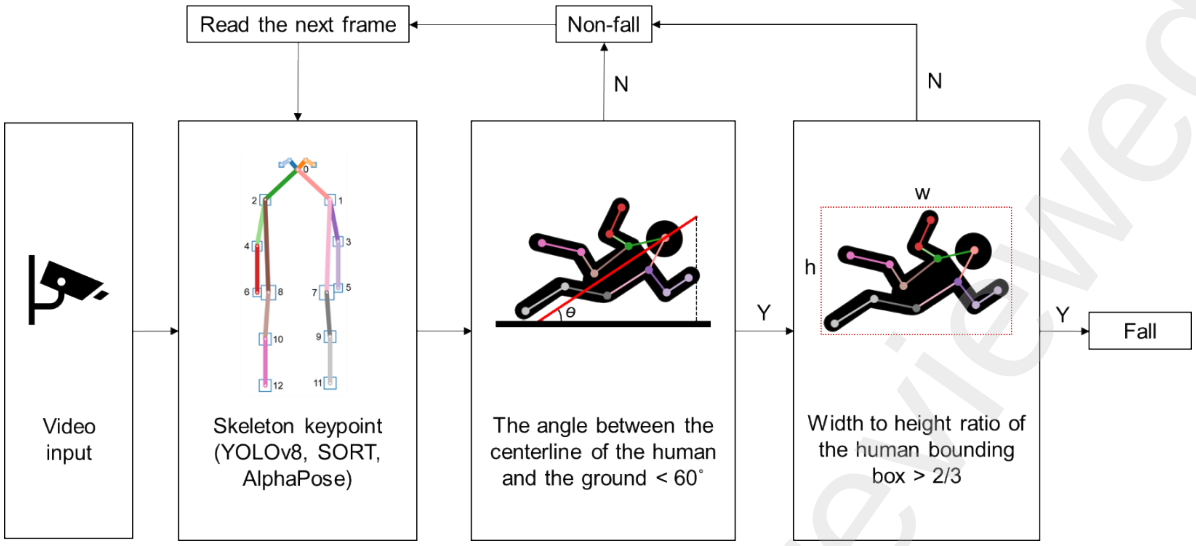


Figure 8. The schematic diagram of threshold-based fall detection model.

We also included a k-Nearest Neighbor (KNN) model, as referenced in (Ramirez et al., 2021) for comparison. Additionally, a Spatial-Temporal Graph Convolutional Network (ST-GCN) (Yan et al., 2018), known for its ability to capture dynamic spatial and temporal patterns in keypoint data was selected due to its structural differences from 1D CNN-LSTM and its proficiency in handling sequential keypoint data, making it a robust candidate for performance benchmarking in fall classification tasks.

As detailed in **Table 3**, the 1D CNN-LSTM model achieved an accuracy of 98.66% for the total dataset, surpassing the threshold-based model by 10.23%, the KNN model by 1.44%, and the ST-GCN model by 0.07%. Additionally, the 1D CNN-LSTM model achieved the highest specificity of 99.10% and maintained a fast inference speed of 6.44 FPS. Even within the Ai-Hub dataset which includes simulated falls in construction sites, the 1D CNN-LSTM model achieved a high accuracy of 97.44%, along with 98.02% sensitivity, and 96.83% specificity. Furthermore, the 1D CNN-LSTM model is more parameter-efficient with only 186,303 parameters compared to 6,141,758 for the ST-GCN model.

In summary, the 1D CNN-LSTM model demonstrates superior performance in accuracy, specificity, and inference speed when compared to other models. This consistent outperformance suggests the robustness of the 1D CNN-LSTM model and its potential for real-time application in various scenarios, including those with complex detection environments as indicated by the Ai-Hub dataset results.

Table 3. Performance comparison between 1D CNN-LSTM and other models.

Dataset	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	FPS
Total	Threshold-based	88.82	67.28	99.09	6.66
	KNN	97.22	94.02	98.74	4.18
	ST-GCN	98.59	97.90	98.82	5.41
	1D CNN-LSTM	98.66	97.32	99.10	6.44
KFall dataset	Threshold-based	91.79	72.49	99.78	
	KNN	98.36	95.49	99.55	
	ST-GCN	99.13	97.94	99.47	

	1D CNN-LSTM	98.97	97.94	99.27
Multiple camera dataset	Threshold-based	70.92	50.60	88.03
	KNN	82.95	75.33	89.37
	ST-GCN	88.39	91.94	86.02
	1D CNN-LSTM	91.61	83.87	96.77
Ai-Hub dataset	Threshold-based	67.89	46.49	95.28
	KNN	93.56	94.91	91.84
	ST-GCN	95.65	99.50	91.53
	1D CNN-LSTM	97.44	98.02	96.83

3.2 Fall classification accuracy according to occlusion level

To rigorously assess the effectiveness of the 1D CNN-LSTM fall classification model, especially under the frequent occlusions characteristic of construction sites, we validate the model's performance with varying levels of occlusion. The rationale behind this specific validation is rooted in the complex nature of construction environments, where workers are often partially or fully occluded by machinery, equipment, or other structures. Accurate fall detection in such scenarios is critical for ensuring worker safety, and it remains a largely unexplored area in fall detection research.

To achieve this, we adopted a novel approach originally developed by Gilroy et.al (Gilroy et al., 2021) for image-based classification of pedestrian occlusion levels. This approach is relevant for our study as it offers a systematic way to quantify occlusion, a key factor in the effectiveness of fall detection at construction sites. This method involves two main steps: keypoint detection and an advanced 2D body surface area estimation. Initially, keypoints for each individual in an image are identified. The visibility of these keypoints is then evaluated - those with confidence scores below a certain threshold are considered occluded. The modified version of the Rule of Nines (Gilroy et al., 2021) is utilized to calculate the visible body surface area, which is then used to determine occlusion levels. As detailed in **Table 4**, occlusion classification is determined by the visibility of associated keypoints.

Table 4. Percentage of total Body Surface Area (BSA) and related keypoints for each semantic body part.

Body Part (% BSA)	Related Keypoints
Head (9%)	Nose or Eyes or Ears
Upper Torso (18%)	L Shoulder + R Shoulder
Upper Left Arm (4.5%)	L Shoulder + L Elbow
Lower Left Arm (4.5%)	L Elbow + L Wrist
Upper Right Arm (4.5%)	R Shoulder + R Elbow
Lower Right Arm (4.5%)	R Elbow + R Wrist
Lower Torso (18%)	L Hip + R Hip
Upper Left Leg (9%)	L Hip + L Knee
Lower Left Leg (9%)	L Knee + L Ankle
Upper Right Leg (9%)	R Hip + R Knee
Lower Right Leg (9%)	R Knee + R Ankle

To evaluate the occlusion levels of the dataset, this systematic method is applied to each fall dataset used in our study. The total dataset shows an occlusion level of 1.94%, the KFall dataset 1.65%, the multiple camera dataset 2.25%, and the Ai-Hub dataset has the highest at 5.65%, likely reflecting the complex environments of construction sites and the presence of many workers. Additionally, occlusion levels are assessed in the context of 'non-fall' and 'fall' scenarios, revealing that falls are associated with a significantly higher occlusion level of 7.06% while non-falls have 0.28% occlusion level, indicating instances of self-occlusion during falls.

In evaluating our 1D CNN-LSTM model's performance against these occlusion levels, we categorized occlusion into 'low' (below 10%) and 'high' (above 10%) as defined in (Li et al., 2016). **Table 5** presents a comprehensive analysis of the fall classification performance across occlusion levels. Remarkably, in terms of accuracy and sensitivity, the model demonstrates consistent performance in both high and low occlusion scenarios. It is noteworthy, however, that specificity tends to decrease in high occlusion situations, primarily due to the lower incidence of non-fall cases in these scenarios. A key aspect of our model's evaluation is the incorporation of confidence scores, which play a vital role in sustaining reliable performance irrespective of occlusion levels. This feature enhances the model's ability to accurately identify falls, even in scenarios where visibility is occluded. The findings from this evaluation highlight the exceptional robustness of our 1D CNN-LSTM model in handling diverse visibility conditions encountered in fall detection. The model's ability to maintain high accuracy and sensitivity across different levels of occlusion signifies its practical applicability in real-world, dynamic environments like construction sites. This adaptability and reliability in occluded conditions mark a significant advancement in the field of fall detection technology.

Table 5. Fall classification performance according to occlusion level.

Occlusion level	Accuracy (%)	Sensitivity (%)	Specificity (%)	Fall ratio(%)	Total window
Low	98.68	96.67	99.19	19.75	5295
High	98.43	99.67	76.47	94.36	319
All	98.66	97.32	99.10	23.99	5611

3.3 Development of a portable AI-based automated safety monitoring system

The development of a portable, AI-based automated safety monitoring system for construction sites addresses the limitations of traditional surveillance methods, such as fixed CCTV cameras that often miss large areas, creating safety blind spots. These conventional systems also suffer from inefficient manual monitoring and increased risk of human errors, alongside challenges like high bandwidth usage, network dependency, and privacy concerns due to continuous video data streaming to external servers.

Our innovative system utilizes edge computing, where an edge device connected to the camera performs internal image analysis, significantly reducing bandwidth use and enhancing privacy by locally processing data. This portable system is particularly advantageous for monitoring high-elevation work areas with sparse worker presence, offering flexibility in coverage where fixed CCTV is inadequate. It autonomously detects fall incidents in real-time, improving the efficiency and responsiveness of safety control teams.

The prototype, as illustrated in **Figure 9**, comprises three main components: the edge device Jetson Xavier NX, a portable monitor, and a camera on a tripod. The Jetson Xavier NX, equipped with NVIDIA's Volta GPU core and 8GB of memory, is central to the system, enabling the deployment of advanced deep learning models for real-time fall detection and response. The portable monitor allows easy interaction with the system, and the camera's versatility in positioning ensures comprehensive site coverage. This portable and adaptive system, shown in **Figure 10**, shows the example of monitoring fall detection by applying our model to video.

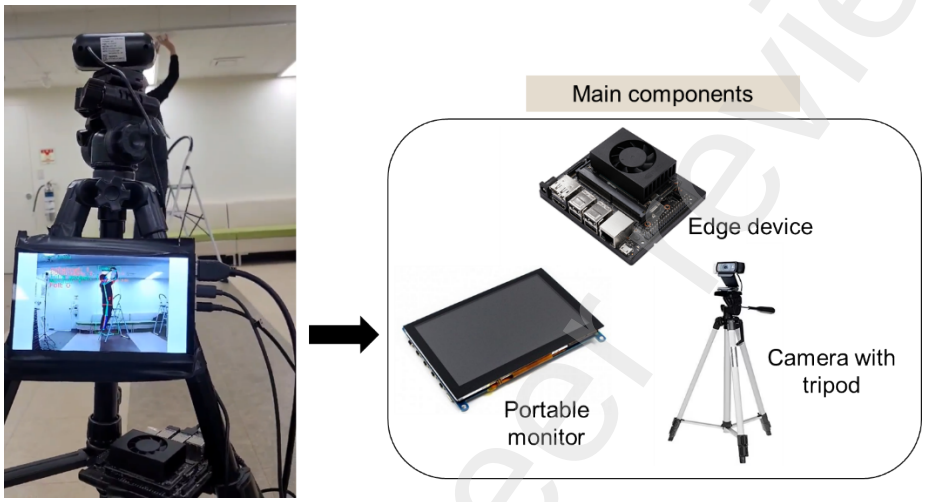


Figure 9. Portable AI-based automated safety monitoring system for fall detection.

(a) Non-fall case (b) Fall case

Figure 10. Examples of monitoring fall status by applying our model to video.

4 Discussion

We have developed a portable, skeleton-based real-time multi-person fall detection model, YOSAP-LSTM, for construction sites using edge devices. The model uses YOLOv8 to detect workers with an accuracy of 93.60%. This model tracks individuals using SORT and AlphaPose for skeleton keypoint analysis, feeding into a 1D CNN-LSTM model that classifies actions as fall incidents with 98.66% accuracy, 97.32% sensitivity, and 99.10% specificity while maintaining a real-time performance of 6.44 FPS. The model also demonstrates adaptability to varied occlusion levels and computational constraints, showcasing the effective

integration of YOLOv8 in enhancing fall detection and the robust deployment capability via edge devices.

It is understandable that while the threshold-based model from the study by Chen et al. (Chen et al., 2020) achieved 97% accuracy, our approach with the broader dataset and varied scenarios shows different results, 88.43% accuracy. The YOSAP-LSTM model, unlike Chen's study which was restricted to simple falling actions, incorporates comprehensive movements such as lying down, significantly broadening its applicability. Furthermore, it is worthwhile to mention that our dataset is substantially larger, comprising 376,006 frames compared to the 100 video actions in Chen's study, enabling more robust and generalizable results. Additionally, the KNN model, as reported by Ramirez et al. (Ramirez et al., 2021), demonstrated a high accuracy of 98.84% when tested with the UP-fall dataset (Martínez-Villaseñor et al., 2019). In our research, the KNN model showed a similar level of effectiveness, achieving an accuracy of 97.22%, underscoring its reliability in diverse fall detection scenarios.

The CNN-LSTM architecture improves the accuracy of time-series data predictions by combining the strengths of CNN and LSTM networks. Time-series data often exhibit nonlinear characteristics, making it challenging for classical statistical methods to forecast accurately (Chandrasekaran & Kumar Paramasivan, 2022b). CNNs and LSTMs are better equipped to handle nonlinearities and process sequences of inputs (Chandrasekaran & Kumar Paramasivan, 2022b; H. J. Kim et al., 2022). Gopala et.al (Gopala et al., 2024) have utilized a CNN-LSTM network to effectively harness temporal and spatial features in sensor data, thereby improving feature extraction and temporal dependency detection in wireless networks. Similarly, Yu et al.'s (X. Yu et al., 2020) ConvLSTM model merges CNN's spatial feature mapping with LSTM's temporal analysis to predict falls more accurately in the elderly using IMU data, outperforming standalone CNN or LSTM models. Our fall detection system combines CNNs and LSTMs to process spatial and temporal information efficiently, achieving similar accuracy and frame rates as the larger ST-GCN model, but with significantly fewer parameters - 186,303 for CNN-LSTM compared to 6,141,758 for ST-GCN - resulting in faster inference and reduced model complexity.

In comparison to existing research, our fall detection model, using the latest improved YOLOv8 on edge devices, outperforms the Nunez et.al study (Núñez-Marcos et al., 2017) which relies on optical flow and VGG-16 with transfer learning, by achieving higher sensitivity and specificity of 97.32% and 99.10%, respectively, against their reported 94.00%. Unlike the Gomes et.al (Gomes et al., 2022) approach, which attains 96.35% accuracy at a processing speed of 25 FPS using a GTX 1060 TI without edge deployment, our model efficiently operates on edge devices. It also surpasses the Nguyen et.al method (Nguyễn Tiến et al., 2023) that integrates YOLOv3-tiny with SORT and AlphaPose, offering higher precision and recall, with a comprehensive dataset of over 323,412 frames compared to their single-dataset usage. Finally, our model demonstrates superior precision, recall, and F1-score metrics over the YOLOv3-based system by Charfi et al. (Charfi et al., 2013), which was tested on a dataset of 220 videos, highlighting our system's efficacy in complex, real-world scenarios.

Change et.al (Chang et al., 2021) presented a fall detection system using the edge device Jetson TX2, achieving a 98.10% accuracy and 10 FPS with OpenPose-light for joint recognition and LSTM for fall classification. However, their model was tested on the Le2i

dataset (Charfi et al., 2013), featuring simulated falls in indoor settings, lacking complex backgrounds and occlusions found in actual construction sites. Lin et.al (B. S. Lin et al., 2022) developed a similar system that applies an improved YOLOv3 and SVM classifier on edge devices, reaching 96% accuracy at 11.5 FPS, but their data was limited to a small set of images from various indoor locations. In contrast, our study advances the field by analyzing a more extensive video dataset, encompassing multi-person fall incidents on construction sites, to fine-tune fall detection for edge device deployment. Furthermore, we have developed a portable, AI-based automated safety monitoring system specifically designed for real-world construction environments. The system's portable configuration, which includes the Jetson Xavier NX edge device, a monitor, and a camera mounted on a tripod, provides extensive coverage and enables autonomous real-time detection of fall incidents. This dramatically enhances the responsiveness of safety control teams, making our system particularly suitable for the dynamic and complex conditions found at construction sites.

Our computer vision-based multi-person fall detection model demonstrates high accuracy but has some limitations. A significant challenge is occlusion, where falls occurring behind obstacles or other individuals might not be captured, leading to false negatives. The model performance is also sensitive to environmental conditions such as poor lighting or extreme weather, potentially affecting its reliability. In real-world application experiments, we have observed that camera placement and the diverse postures of workers can influence the accuracy of the model. For example, when workers are positioned laterally, at a distance, or near the camera's periphery, the model's effectiveness diminishes. Fang et.al (C. Fang et al., 2022) observed that complex postures, such as bending or squatting, increase false alarms in fall detection systems, recommending side-placed cameras in a monocular vision system for improved accuracy. Ponce et.al (Ponce et al., 2020) also found that while ceiling-mounted cameras reduce occlusion, they are less effective at capturing essential vertical motions, with lateral camera positioning proving to be more effective than frontal placement for fall detection. To mitigate these issues, adopting a multi-view strategy with two cameras set orthogonally to the subject's motion has been theorized to significantly improve detection rates (Thome et al., 2008). Espinosa et al. (Espinosa et al., 2019) demonstrated that using a VGG-16 CNN architecture with dual-camera views maintains effective performance, even when one camera is obstructed, thereby ensuring the reliability of fall detection. Future enhancements to our model could include integrating additional sensors, such as thermal imaging, to reduce dependence on visual data. This approach could help counteract challenges posed by inadequate lighting and visual obstructions. Enriching the dataset with a broader range of fall scenarios and environmental conditions is expected to further strengthen the model's robustness and accuracy. An additional research focus could be enhancing the model's scalability and efficiency for broader applications across varied construction sites, each with unique safety and computational needs. Last but not least, leveraging recent advancements in edge computing hardware could augment the model's capabilities and facilitate the execution of more complex computational tasks in real-time on-site.

5 Conclusion

We have successfully developed YOSAP-LSTM, a practical and robust real-time skeleton-

based multi-person fall detection model, aiming to enhance worker safety on construction sites utilizing edge devices. Our model is built on a comprehensive fall video dataset comprising 376,006 frames from diverse environments, including realistic construction fall scenarios from the Ai-Hub dataset. Integrating YOLOv8, our model achieves an impressive 93.60% accuracy in worker detection. The incorporation of SORT and AlphaPose ensures precise extraction of skeleton keypoints, which are subsequently analyzed by a 1D CNN-LSTM model, classifying falls with remarkable overall accuracy—98.66%, 97.32% sensitivity, and 99.10% specificity, operating at 6.44 FPS on the edge device Jetson Xavier NX. Notably, our model maintains a high accuracy of 97.44% for fall classification within the challenging Ai-Hub dataset, underscoring its adaptability to complex site environments. Even in varying levels of occlusion, the model exhibits robust performance, showcasing its reliability. By enabling safety managers to conduct remote surveillance capabilities, our model ensures swift responses to fall incidents and helps to enhance on-site safety.

Declaration of interest statement

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.

Acknowledgments

This work was supported by Samsung Display Co., Ltd (G01220224) and the Basic Science Research Program of the National Research Foundation of Korea, Ministry of Science and ICT, (NRF2022R1F1A1061045, NRF-2022M3J6A1063021).

CRediT authorship contribution statement

Author 1: Data curation, Methodology, Software, Formal analysis, Writing – Original draft.

Author 2: Methodology, Formal analysis.

Author 3 (corresponding author): Conceptualization, Funding acquisition, Project administration, Writing – Review & Editing.

References

Abbate, S., Avvenuti, M., Bonatesta, F., Cola, G., Corsini, P., & Vecchio, A. (2012). A smartphone-based fall detection system. *Pervasive and Mobile Computing*, *8*(6), 883–899. https://doi.org/10.1016/j.pmcj.2012.08.003

Adhikari, K., Bouchachia, H., & Nait-Charif, H. (2017). Activity recognition for indoor fall detection using convolutional neural network. *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, 81–84.

Agarwal, T. M., Muneer, M., Asim, M., Awad, M., Afzal, Y., Al-Thani, H., Alhassan, A., Mollazehi, M., & El-Menyar, A. (2020). Psychological trauma in different mechanisms

of traumatic injury: A hospital-based cross-sectional study. *PLoS ONE*, *15*(11 November). https://doi.org/10.1371/journal.pone.0242849

AI-HUB Dataset. (2020). *AI-HUB Dataset, https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=163*.

An, J., Kim, J., Lee, H., Kim, J., Kang, J., Shin, S., Kim, M., Hong, D., & Woo, S. S. (2021). VFP290k: A large-scale benchmark dataset for vision-based fallen person detection. *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Auvinet, E., Rougier, C., Meunier, J., St-Arnaud, A., & Rousseau, J. (2010). *Multiple cameras fall data set*.

Bai Z Y, Ding QC, Xu HL & Wu CD. (2023). Human similar action recognition by fusing saliency image semantic features. *Journal of Image and Graphics*, 28(09): 2872-2886. In Chinese.

Baldewijns, G., Debard, G., Mertes, G., Vanrumste, B., & Croonenborghs, T. (2016). Bridging the gap between real-life data and simulated data by providing a highly realistic Fall dataset for evaluating camera-based fall detection algorithms. *Healthcare Technology Letters*, *3*(1), 6–11. https://doi.org/10.1049/htl.2015.0047

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. *ArXiv Preprint ArXiv:2006.10204*.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). *Simple Online and Realtime Tracking*. https://doi.org/10.1109/ICIP.2016.7533003

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.

Castellano, P. S. (2023). AXONOMÍA DE LAS GARANTÍAS JURÍDICAS EN EL EMPLEO DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL. *Revista de Derecho Politico*, *117*, 153–196. https://doi.org/10.13039/501100011033

Chan, A. P. C., Wong, F. K. W., Chan, D. W. M., Yam, M. C. H., Kwok, A. W. K., Lam, E. W. M., & Cheung, E. (2008). Work at Height Fatalities in the Repair, Maintenance, Alteration, and Addition Works. *Journal of Construction Engineering and Management*, *134*(7), 527–535. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:7(527)/ASSET/2DA6E9F4-6F2B-4C55-9556-5CFAEC4831C2/ASSETS/IMAGES/LARGE/8.JPG

Chandrasekaran, R., & Kumar Paramasivan, S. (2022a). A State-of-the-Art Review of Time Series Forecasting Using Deep Learning Approaches. *International Journal on Recent and Innovation Trends in Computing and Communication*, *10*(12), 92–105. https://doi.org/10.17762/ijritcc.v10i12.5890

Chandrasekaran, R., & Kumar Paramasivan, S. (2022b). A State-of-the-Art Review of Time

Series Forecasting Using Deep Learning Approaches. *International Journal on Recent and Innovation Trends in Computing and Communication*, *10*(12), 92–105. https://doi.org/10.17762/ijritcc.v10i12.5890

Chang, W. J., Hsu, C. H., & Chen, L. B. (2021). A Pose Estimation-Based Fall Detection Methodology Using Artificial Intelligence Edge Computing. *IEEE Access*, *9*, 129965–129976. https://doi.org/10.1109/ACCESS.2021.3113824

Charfi, I., Miteran, J., Dubois, J., Atri, M., & Tourki, R. (2013). Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification. *Journal of Electronic Imaging*, *22*(4), 041106. https://doi.org/10.1117/1.jei.22.4.041106

Chen, W., Jiang, Z., Guo, H., & Ni, X. (2020). Fall Detection based on key points of human-skeleton using openpose. *Symmetry*, *12*(5). https://doi.org/10.3390/SYM12050744

Choi, S. D., Guo, L., Kim, J., & Xiong, S. (2019). Comparison of fatal occupational injuries in construction industry in the United States, South Korea, and China. *International Journal of Industrial Ergonomics*, *71*, 64–74. https://doi.org/10.1016/j.ergon.2019.02.011

Du, Y., Fu, Y., & Wang, L. (2015). Skeleton based action recognition with convolutional neural network. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 579–583.

Espinosa, R., Ponce, H., Gutiérrez, S., Martínez-Villaseñor, L., Brieva, J., & Moya-Albor, E. (2019). A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset. *Computers in Biology and Medicine*, *115*. https://doi.org/10.1016/j.compbiomed.2019.103520

Fanca, A., Puscasiu, A., Gota, D.-I., & Valean, H. (2019). Methods to minimize false detection in accidental fall warning systems. *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*, 851–855.

Fang, C., Xiang, H., Leng, C., Chen, J., & Yu, Q. (2022). Research on Real-Time Detection of Safety Harness Wearing of Workshop Personnel Based on YOLOv5 and OpenPose. *Sustainability (Switzerland)*, *14*(10). https://doi.org/10.3390/su14105872

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., & Lu, C. (2022). *AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time*. http://arxiv.org/abs/2211.03375

Feng, Q., Gao, C., Wang, L., Zhao, Y., Song, T., & Li, Q. (2020). Spatio-temporal fall event detection in complex scenes using attention guided LSTM. *Pattern Recognition Letters*, *130*, 242–249. https://doi.org/10.1016/j.patrec.2018.08.031

Gallo, G., Rienzo, F. Di, Garzelli, F., Ducange, P., & Vallati, C. (2022). A Smart System for Personal Protective Equipment Detection in Industrial Environments Based on Deep Learning at the Edge. *IEEE Access*, *10*, 110862–110878. https://doi.org/10.1109/ACCESS.2022.3215148

Gilroy, S., Glavin, M., Jones, E., & Mullins, D. (2021). Pedestrian Occlusion Level

Classification using Keypoint Detection and 2D Body Surface Area Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, *2021-October*, 3826–3832. https://doi.org/10.1109/ICCVW54120.2021.00427

Gomes, M. E. N., Macêdo, D., Zanchettin, C., de-Mattos-Neto, P. S. G., & Oliveira, A. (2022). Multi-human Fall Detection and Localization in Videos. *Computer Vision and Image Understanding*, *220*. https://doi.org/10.1016/j.cviu.2022.103442

Gopala, T., Raviram, V., & NL, U. K. (2024). Detecting Security Threats in Wireless Sensor Networks using Hybrid Network of CNNs and Long Short-Term Memory. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(1s), 704–722.

Guerrero, J. C. E., España, E. M., Añasco, M. M., & Lopera, J. E. P. (2022). Dataset for human fall recognition in an uncontrolled environment. *Data in Brief*, *45*. https://doi.org/10.1016/j.dib.2022.108610

Hu, K., Rahmandad, H., Smith-Jackson, T., & Winchester, W. (2011). Factors influencing the risk of falls in the constructionindustry: A review of the evidence. *Construction Management and Economics*, *29*(4), 397–416. https://doi.org/10.1080/01446193.2011.558104

Inturi, A. R., Manikandan, V. M., & Garrapally, V. (2023). A Novel Vision-Based Fall Detection Scheme Using Keypoints of Human Skeleton with Long Short-Term Memory Network. *Arabian Journal for Science and Engineering*, *48*(2), 1143–1155. https://doi.org/10.1007/s13369-022-06684-x

Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics*. https://github.com/ultralytics/ultralytics

Kim, H. J., Depoian, A. C., Bailey, C. P., & Guturu, P. (2022). *Novel neural network architecture for energy prediction*. 5. https://doi.org/10.1117/12.2619143

Kim, Y., Jung, H., Koo, B., Kim, J., Kim, T., & Nam, Y. (2020). Detection of pre-impact falls from heights using an inertial measurement unit sensor. *Sensors (Switzerland)*, *20*(18), 1–14. https://doi.org/10.3390/s20185388

Kwolek, B., & Kepski, M. (2014). Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, *117*(3), 489–501. https://doi.org/10.1016/j.cmpb.2014.09.005

Lee, J., & Hwang, K. il. (2022). YOLO with adaptive frame control for real-time object detection applications. *Multimedia Tools and Applications*, *81*(25), 36375–36396. https://doi.org/10.1007/s11042-021-11480-0

Leite, G., Silva, G., & Pedrini, H. (2019). Fall detection in video sequences based on a three-stream convolutional neural network. *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 191–195. https://doi.org/10.1109/ICMLA.2019.00037

Lin, B. S., Yu, T., Peng, C. W., Lin, C. H., Hsu, H. K., Lee, I. J., & Zhang, Z. (2022). Fall

Detection System with Artificial Intelligence-Based Edge Computing. *IEEE Access*, *10*, 4328–4339. https://doi.org/10.1109/ACCESS.2021.3140164

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-October*, 2999–3007. https://doi.org/10.1109/ICCV.2017.324

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., & Kot, A. C. (2019). *NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding*. https://doi.org/10.1109/TPAMI.2019.2916873

Li, X., Flohr, F., Yang, Y., Xiong, H., Braun, M., Pan, S., Li, K., & Gavrila, D. M. (2016). A new benchmark for vision-based cyclist detection. *2016 IEEE Intelligent Vehicles Symposium (IV)*, 1028–1033.

Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., & Peñafort-Asturiano, C. (2019). Up-fall detection dataset: A multimodal approach. *Sensors (Switzerland)*, *19*(9). https://doi.org/10.3390/s19091988

McKinley, S., & Levine, M. (1998). Cubic spline interpolation. *College of the Redwoods*, *45*(1), 1049–1060.

Moutsis, S. N., Tsintotas, K. A., Kansizoglou, I., An, S., Aloimonos, Y., & Gasteratos, A. (2023). Fall detection paradigm for embedded devices based on YOLOv8. *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–6. https://doi.org/10.1109/IST59124.2023.10355696

Nath, N. D., Behzadan, A. H., & Paal, S. G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, *112*, 103085. https://doi.org/10.1016/J.AUTCON.2020.103085

Newaz, N. T., & Hanada, E. (2023). The Methods of Fall Detection: A Literature Review. In *Sensors* (Vol. 23, Issue 11). MDPI. https://doi.org/10.3390/s23115212

Nguyen, H. H., Ta, T. N., Nguyen, N. C., Bui, V. T., Pham, H. M., & Nguyen, D. M. (2021). YOLO Based Real-Time Human Detection for Smart Video Surveillance at the Edge. *ICCE 2020 - 2020 IEEE 8th International Conference on Communications and Electronics*, 439–444. https://doi.org/10.1109/ICCE48956.2021.9352144

Nguyễn Tiến, Đ., Đỗ Hoàng, V., & Nguyễn Văn, T. (2023). *Vision-based fall detection system for the elderly using image processing and deep learning*. 22. https://doi.org/10.1117/12.2670053

Núñez-Marcos, A., Azkune, G., & Arganda-Carreras, I. (2017). Vision-based fall detection with convolutional neural networks. *Wireless Communications and Mobile Computing*, *2017*. https://doi.org/10.1155/2017/9474806

Pal, A., & Hsieh, S.-H. (2020). Vision based construction site monitoring: a review from construction management point of view. *Enabling the Development and Implementation of Digital Twins: Proceedings of the 20th International Conference on Construction Applications of Virtual Reality, Teesside University*, 44–55.

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517. https://doi.org/10.1016/J.AUTCON.2020.103517

Ponce, H., Martínez-Villaseñor, L., & Nuñez-Martínez, J. (2020). Sensor location analysis and minimal deployment for fall detection system. *IEEE Access*, *8*, 166678–166691. https://doi.org/10.1109/ACCESS.2020.3022971

Ramirez, H., Velastin, S. A., Fabregas, E., Meza, I., Makris, D., & Farias, G. (2021). *Fall detection using human skeleton features*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*. http://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Roboflow, H. H. W. D. (2022). *Roboflow, Hard Hat Workers Dataset, https://public.roboflow.com/object-detection/hard-hat-workers*.

Rouali, M. L., Boulahia, S. Y., & Amamra, A. (2023). Structure and Sequencing Preserving Representations for Skeleton-based Action Recognition Relying on Attention Mechanisms. *Journal of Signal Processing Systems*. https://doi.org/10.1007/s11265-023-01892-6

Sarkar, S., Chang, C. K., Wang, X., Ellul, J., & Azzopardi, G. (2020). Elderly Fall Detection Systems: A Literature Survey. *Frontiers in Robotics and AI | Www.Frontiersin.Org*, *1*, 71. https://doi.org/10.3389/frobt.2020.00071

Schröter, E., Thanh Nghi, D., & Schneider, A. (2023). Development of an Intelligent Walking Aid for Fall Detection. *Current Directions in Biomedical Engineering*, *9*(1), 287–290. https://doi.org/10.1515/cdbme-2023-1072

Seo, J., Han, S., Lee, S., & Kim, H. (2015). Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, *29*(2), 239–251. https://doi.org/10.1016/J.AEI.2015.02.001

South Korea Ministry of Employment and Labor. (2020). *South Korea Ministry of Employment and Labor. (2020). Industrial Accident Investigation Overview. Retrieved May 10, 2023, from https://www.moel.go.kr/info/publict/publictDataView.do?bbs_seq=20211201900*.

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks*. http://arxiv.org/abs/1909.09586

Sucerquia, A., David López, J., & Vargas-Bonilla, J. F. (2017). *SisFall: A Fall and Movement Dataset*. https://doi.org/10.3390/s17010198

Thome, N., Miguet, S., & Ambellouis, S. (2008). A real-time, multiview fall detection system: A LHMM-based approach. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1522–1532. https://doi.org/10.1109/TCSVT.2008.2005606

Wang, Z., Wu, Y., Yang, L., Thirunavukarasu, A., Evison, C., & Zhao, Y. (2021). Fast personal protective equipment detection for real construction sites using deep learning approaches. *Sensors*, *21*(10). https://doi.org/10.3390/s21103478

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. *2017 IEEE International Conference on Image Processing (ICIP)*, 3645–3649.

Yan, S., Xiong, Y., & Lin, D. (2018). *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition*. http://arxiv.org/abs/1801.07455

Yonghye Kwon. (2019). *Y. Kwon, "Yolo_Label", Github, Available online: https://github.com/developer0hye/Yolo_Label*. https://github.com/developer0hye/Yolo_Label?tab=readme-ov-file

Yu, M., Gong, L., & Kollias, S. (2017). *Computer Vision Based Fall Detection by a Convolutional Neural Network*. https://doi.org/10.1145/3136755.3136802

Yu, X., Jang, J., & Xiong, S. (2021). A Large-Scale Open Motion Dataset (KFall) and Benchmark Algorithms for Detecting Pre-impact Fall of the Elderly Using Wearable Inertial Sensors. *Frontiers in Aging Neuroscience*, *13*. https://doi.org/10.3389/fnagi.2021.692865

Yu, X., Qiu, H., & Xiong, S. (2020). A Novel Hybrid Deep Neural Network to Predict Pre-impact Fall for Older People Based on Wearable Inertial Sensors. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.00063

Zhang, Y., Gan, J., Zhao, Z., Chen, J., Chen, X., Diao, Y., & Tu, S. (2023). A real-time fall detection model based on BlazePose and improved ST-GCN. *Journal of Real-Time Image Processing*, *20*(6). https://doi.org/10.1007/s11554-023-01377-6

Zhang, Z., Conly, C., & Athitsos, V. (2015, July 1). A survey on vision-based fall detection. *8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2015 - Proceedings*. https://doi.org/10.1145/2769493.2769540

Zheng, H., & Liu, Y. (2022). Lightweight Fall Detection Algorithm Based on AlphaPose Optimization Model and ST-GCN. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/9962666

Zheng, Y., Zhang, D., Yang, L., & Zhou, Z. (2019). Fall detection and recognition based on gcn and 2d pose. *2019 6th International Conference on Systems and Informatics (ICSAI)*, 558–562.