

محمد باعطيہ

Construction V 1.12.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:406736281

Submission Date

Aug 11, 2026, 6:50 PM GMT+3

Download Date

Aug 14, 2026, 5:39 PM GMT+3

File Name

Construction V 1.12.docx

File Size

6.2 MB

35 Pages

7,835 Words

44,840 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



AI-Based Construction Site Safety Monitoring System

Prepared By

Name: Mohammed Ali Alhawshabi

ID No.: 261002622

Name: Mohammed Omar Baatiyyah

ID No.: 261002493

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Mohamed Eassa

College of Informatics, Midocean University, Comoros

Email: 261002493@st.midocean.edu.km

Email: 261002622@st.midocean.edu.km

Acknowledgements

We would like to thank our supervisor, Dr. Mohamed Essa, for his continuous guidance and valuable feedback throughout this research. His expertise in both artificial intelligence and Construction sites has greatly influenced the direction of this work. His comments challenged us to think more critically and approach the research from a scientific perspective rather than a purely technical one.

We are also thankful to the faculty members at the College of Informatics, Midocean University, for providing a supportive academic environment that inspired independent thinking and research development.

Finally, we would like to express our appreciation to our families for their patience, encouragement, and support during this journey.

Abstract

As it is known that construction sites are inherently dangerous, containing numerous sharp tools that can lead to injuries of varying severity. Preventing such injuries is often difficult due to a multitude of factors, including the failure to wear protective equipment on construction sites. One of the most important methods for monitoring and ensuring workers on construction sites is a visual inspection by the site supervisor to verify that all workers are wearing appropriate protective gear. Checking the use of protective gear at construction sites is extremely important and requires significant human intervention, not only when workers enter the site but also during their daily work hours.

To ensure worker protection, the system uses artificial intelligence to check whether workers are wearing adequate protective gear. Additionally, it monitors worker movements, and in the event of an injury, it immediately alerts paramedics to expedite the rescue operation.

In large and crowded construction sites, it is difficult to monitor everyone at the same time, which may require more personnel for monitoring, thus increasing the cost. In addition, artificial intelligence helps in detecting accidents immediately and quickly, which contributes to saving lives more and faster. This study proves that the machine is not just a means of sorting files but can also effectively prevent fatalities by shifting monitoring from a reactive to a proactive approach as much as possible, where achieving zero dangerous incidents is the primary goal.

Contents

ACKNOWLEDGEMENTS	II
ABSTRACT	III
LIST OF ABBREVIATIONS	V
LIST OF FIGURES	VI
LIST OF TABLES	VII
CHAPTER 1: INTRODUCTION	1
1.1 OVERVIEW.....	1
1.2 PROBLEM DEFINITION	2
1.3 THESIS OBJECTIVES.....	4
1.3.1 Safety Monitoring.....	4
1.3.2 Injury and Fall Detection.....	4
1.4 THESIS CONTRIBUTION	5
1.4.1 Unified dual architecture:	5
1.4.2 Aware Hazard Identification	5
1.4.3 Comprehensive Fall and Injury Detection.....	5
1.4.4 A Scalable Template for the Construction Industry	5
1.5 THESIS ORGANIZATION	6
CHAPTER 2: BACKGROUND AND RELATED WORK	7
2.1 INTRODUCTION	7
2.2 BACKGROUND	8
2.2.1 Computer Vision.....	11
2.2.2 Deep Learning	12
2.3 RELATED WORK	18
2.3.1 Safety Monitoring and PPE Detection.....	18
2.3.2 Injury and Fall Detection & Prediction	20
REFERENCES	27

List of Abbreviations

- AI: Artificial Intelligence.
- CNN: Convolutional Neural Network.
- DL: Deep Learning.
- ML: Machine Learning.
- F1-score: Harmonic mean of precision and recall.
- ViT: Vision Transformer.
- CvT: Convolutional Vision Transformer.

List of Figures

Figure 1.1: Example of Body PPE need to be worn on construction site .**Error!**

Bookmark not defined.

Figure 2.1: PPE needs to be worn on construction site7

Figure 2.2: showing how AI can detect and analyze the CCTV data9

Figure 2.3: Schematic overview of a deep learning flow for construction safety monitoring10

Figure 2.4: Visualizing Intersection over Union (IoU) scores12

Figure 2.5: Schematic overview of a deep learning flow for construction safety monitoring15

Figure 2.6: Mapping human keypoints for pose estimation17

List of Tables

Table 2.1: Summary of Related Work on Construction Site Safety Monitoring

Chapter 1: Introduction

1.1 Overview

Safety generally relies on visual observation, with supervisors moving around construction sites to monitor workers' behaviour, including any hazardous actions. This method, known as behavioural screening, is useful but faces several challenges in practice. It requires a large team of observers who often rely solely on their understanding of behaviour on construction sites [1][2]. Simply, the human eye cannot see everything around it. This is not a shortcoming of the work or effort, but the sheer size of the construction site must be considered.

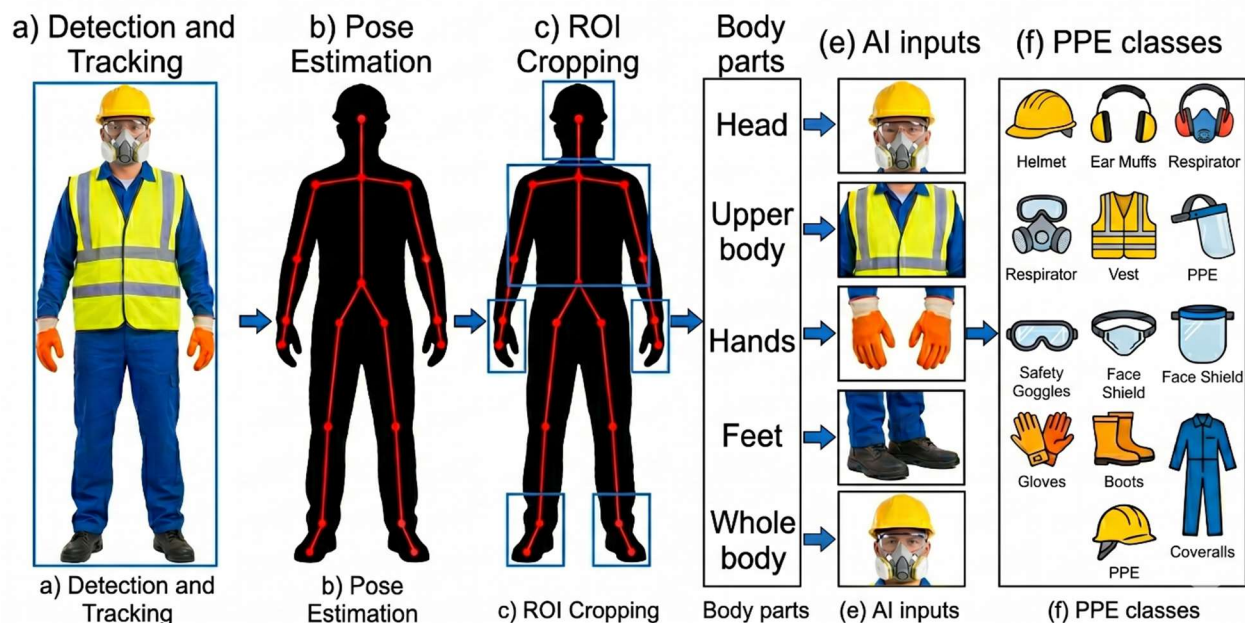


Figure 1.1: Example of Body PPE needs to be worn on construction site

The human eye often misidentifies problems, and relying on manual supervision can create safety gaps that human observers may not be able to detect. This is why engineers rely on computer vision technologies to interpret and process closed-circuit television cameras CCTV footage from construction sites.

By enabling machines to analyse and sort CCTV footage, companies can have a monitor on construction sites 24/7 without fatigue or boredom (even in low light).

This eliminates the need for intermittent searches by human supervisors and does not impede workflow or speed.

Artificial intelligence can detect, analyses, and identify individuals who do not wear protective gear such as body Armor, gloves, and impact resistant footwear, as well as determine the appropriate equipment for each work environment. in additions, AI can predict whether a worker on-site has been fallen, potentially to faster rescue times compared to human intervention. [1] [8] [14].

However, PPE alone is not enough to protect employees but proactively checking whether a worker is wearing appropriate PPE helps significantly reduce workplace injuries on construction sites and save a life [1].

1.2 Problem Definition

Construction sites are among the most dangerous places where workers are exposed to serious injuries and even fatalities, such as falls from heights or collisions with heavy or dangerous equipment. Most studies indicate that injuries on construction sites result from unsafe actions and behaviors by workers. Therefore, risk management at these sites is extremely important and complex due to the constantly changing behaviors of workers during their shift [15].

To protect construction workers, construction companies require strict guidelines for the use of personal protective equipment (PPE) such as safety vests, goggles, and high-visibility vests. However, ensuring workers consistently wear PPE is a significant challenge, as many disregard safety rules due to discomfort, forgetfulness, or a lack of constant supervision, potentially fast to serious injuries.

Currently, safety rules are applied at construction sites by field safety supervisors. Studies have shown that such an approach is ineffective and has many human flaws because it relies on the human element, which may make mistakes and may not be available at all times when workers are present. Therefore, it is difficult for them to monitor everyone, especially in large and active construction sites. As a result, incorrect practices and violations of safety rules at construction sites may not be discovered.

There is currently a gap in technological solutions that could be better developed through computer vision systems. Some existing systems may be able to

determine if a worker is wearing a helmet (stationary objects), others only track the movement of objects, and a few can analyze the worker's dynamic movements. Simply tracking the worker's hand movements is insufficient, it's necessary to know if they are handling a sharp tool and require specific safety systems for such tasks [1] [8].

The system proposed in this study will resolve two main safety issues in construction sites: The first concerns workers' adherence to the use of various types of safety clothing and evaluate if appropriate clothing is being worn for the type of machinery being operated. The second issue related to detecting potential worker falls during workers' daily tasks by analyzing skeletal movement patterns.

Accordingly, the scope of the proposed system will be divided into:

First: Safety monitoring, where the focus will be on ensuring that workers at construction sites wear personal protective equipment (PPE), and linking the type of work to the type of protection used, to safeguard workers.

Second: fall detection, where the focus will be on analyzing workers' postures and expected physical movements that may lead to potential injuries or dangerous falls, thus supporting rapid detection and immediate response.

1.3 Thesis Objectives

The main objective of this research is to develop a computer vision based intelligent camera monitoring system for improved safety in construction sites. The objectives are divided into two categories:

1.3.1 Safety Monitoring

- To design and train a computer vision model to accurately detect safety helmets and high visibility vests in real-time, based on reliable image data sets.
- To detect dangerous tools, machinery, and construction equipment at construction sites
- To identify unsafe worker movements while handling equipment of various types at construction sites

1.3.2 Injury and Fall Detection

- To develop a system that analyzes worker movements through skeletal movements and determines the probability of a worker falling before it happens.
- To assess the interaction of construction workers with the environment (man-machine/material) to identify unsafe behaviors before they cause accidents.
- To identify potential injuries at construction sites based on body position to support rapid response

In addition, the system will be evaluated using appropriate performance metrics for each task, like precision, recall, F1-score, mean Average Precision (mAP) for object detection, and task specific accuracy where applicable, and inference speed under different environmental conditions.

1.4 Thesis Contribution

Based on previous studies and research, this research offers a practical contribution to the field of safety at construction sites:

1.4.1 Unified dual architecture: Most current studies present two separate models for detecting safety devices and detecting object movements, but in our study we propose a single system for both models to process multiple layers of data simultaneously, which helps reduce processing redundancy and provides a more efficient monitoring solution.

1.4.2 Aware Hazard Identification: In this study, we will develop the concept of integrating machine with human movement. The study will not only determine whether the worker is wearing appropriate protective clothing but will also identify if they are near a sharp machine without wearing the appropriate protective clothing, as well as their body position relative to the machine. This will help reduce risks or prevent accidents altogether.

1.4.3 Comprehensive Fall and Injury Detection: While most current studies focus on safety system violations, we will examine continuous and sudden body movements, utilizing surveillance video footage to help determine if there is a possibility of a worker falling for any reason and sending medical assistance immediately.

1.4.4 A Scalable Template for the Construction Industry: This thesis closes the gap between theoretical AI models and practical safety management, by providing a highly scalable and cost-effective template. This new system changes the industry from reactive safety management (after the accident) to proactive prevention (before the injury) and provides a realistic alternative to manual safety inspections.

1.5 Thesis Organization

The rest of this thesis is organized as follows.

Chapter 2 discusses the theoretical background and relevant research areas.

Chapter 3 presents the proposed methodology.

Chapter 4 analyzes the experimental results.

Chapter 5 concludes the study and suggests directions for future work.

Chapter 2: Background and Related Work

2.1 Introduction

Safety procedures at construction sites began in 1970 by Occupational Safety and Health Administration (OSHA) organization as the construction industry experienced rapid growth and development, safety and protection measures were established at that time because of increasing of construction injuries on world, and based on this, safety was divided into physical safety, operational guidelines, and basic protective equipment. All these standards focus on the safety and health of workers at construction sites, protecting their bodies and lives from injuries, thus emphasizing that human safety is the fundamental requirement [16].

The typical safety hierarchy at construction sites can be further illustrated in Figure 2.1, which shows the relationship between hazard removal, safety controls, and the last line of defense. Later, safety authorities developed such (OSHA) further developed these frameworks by updating rigorous compliance checklists and implementing penalties for safety violations [16].

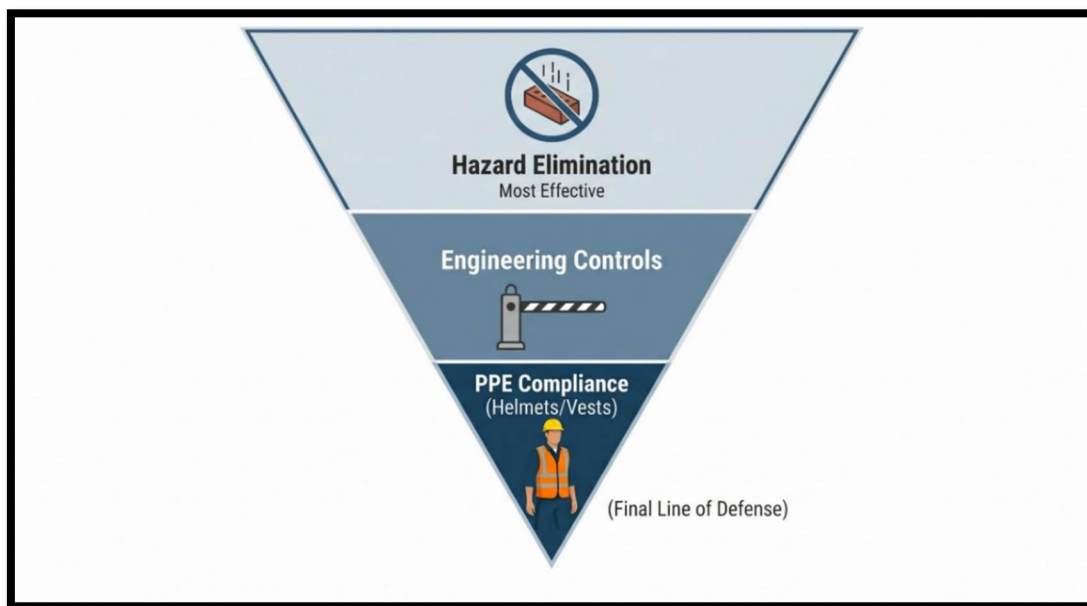


Figure 2.1: PPE needs to be worn on construction site

With the safety systems recommended by international safety organizations for construction sites, safety checks were previously conducted by on site supervisors who would tour construction facilities and visually inspect the availability of safety systems. While these systems have been in use for years, this method remains somewhat effective, it has become impractical due to the sheer volume of work and the scale of construction projects in our time [15].

2.2 Background

Currently, with the increasing growth of construction projects and the growing number of workers and their movement within these projects, safety problems have begun to emerge. These massive sites and industrial cities have begun to generate enormous volumes of work that are difficult to monitor manually effectively [15]. As an alternative solution, the collective safety system was launched, enabling construction contractors to report hazards immediately upon their occurrence. This demonstrates how collective human cooperation contributes to reducing risks.

Because of that, the monitoring of construction sites has continued to expand with the addition of (CCTV) and drones for wide area surveillance. For this reason, the demand for automated safety monitoring has become a matter of interest to researchers in the field of artificial intelligence [15].

From this perspective, the roles of artificial intelligence, machine learning, and deep learning in image and video analysis have become prominent. AI algorithms can analyze, detect, and improve hazard detection at construction sites [15]. Conversely, deep learning can use neural networks to automatically extract complex visual movements and actions. Both methods can be very useful in addressing the problem of adhering to safety standards at construction sites, given that construction site images contain complex visual information that the human eye cannot manually infer, process, and encode. This is evident in the visual contrast between field supervisors when making decisions regarding the implementation of safety systems, as illustrated in the figure 2.2.



Figure 2.2: showing how AI can detect and analyze the CCTV data

The figure 2.2 illustrates examples of construction site images from the safety category, which visual characteristics such as dust, lighting, and other factors. Detecting safety violations under these conditions presents a big challenge for automated detection through image and video analysis. The different positions and interactions of objects or humans make deep learning techniques highly effective for directly identifying safety violations from detailed image Pixels [1][3].

From a computer vision perspective, site safety detection is a real-world problem for identifying and classifying objects. This study aims to detect different types of safety equipment from classified images used at construction sites. Deep learning systems, such as Convolutional neural networks (CNN), Region based Convolutional Neural Network (R-CNN) using Single Stage Detectors (SSD) that allow the model to detect the humane wear PPE by balancing the speed and accuracy to achieve this goal in this case [1][3].

Before the deep learning in artificial intelligence, machine learning was used for basic functions such as detecting the color of protective vests, identifying the shapes of protective equipment, and so on using classical machine learning classifiers and use Support Vector Machines (SVM) with Histogram of Oriented Gradients

(HOG) by breaking image into grille and direction. Also using Random Forests which depend on Yes or NO to define the color or using k-Nearest Neighbours (k-NN) by comparing the new image with labeled on database. With these advanced techniques, they faced challenges in determining all outputs from (CCTV) due to environmental conditions and certain limitations in construction sites [1][14].

With the growing interest in neural networks (CNN) in industrial automation, safety results have shown remarkable success by developed an integrated framework pairing YOLO (for precise object and safety sign detection) with ST-GCN (for skeleton-based human motion capture) to evaluate real-time situational risks and compliance detection accuracy increased from 61.61% to an impressive 99.11% [2].

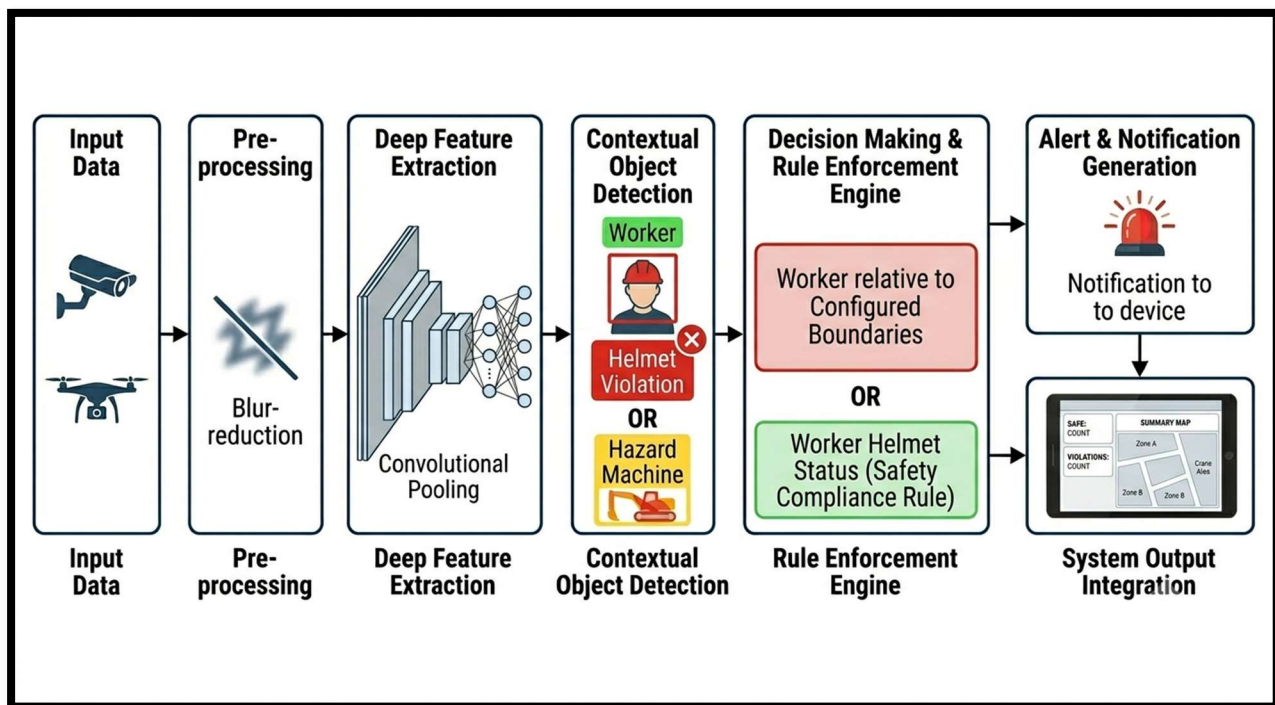


Figure 2.3: Schematic overview of a deep learning flow for construction safety monitoring

The figure 2.3 shows how images are processed, and how they are used to identify hazards and potential safety issues on construction sites. It starts with the camera taking the picture and ends with the identification of the expected safety conditions and their corresponding alerts. This example demonstrates how deep learning can be leveraged to extract and utilize data.

In addition to neural network techniques (CNN), other technologies have been discovered in the field of construction safety to detect more the only static objects

are image and skeleton based parameterized approaches, which deep learning find the skeleton to perform real time worker identity for unsafe behaviors and hazard [1]. Another technique called (Vision Transformers) which can help the model better and more detect and recognize construction sites, such as identifying the relationship between a worker and a heavy crane operating nearby, example the moment a worker holds a ladder and places one foot on the first step is identified as the optimal moment to intercept and warn the worker before they climb too high. In some studies, semi-supervised learning was used, combining some labeled safety datasets and unlabeled continuous (CCTV) footage to reduce reliance on manual classification. Our conclusion here is that current research in this field is gradually and broadly expanding to address complex construction environments [4].

This research combines artificial intelligence (AI) and computer vision to develop an automated system that helps protect construction workers from safety hazards and predicts accidents before they occur.

2.2.1 Computer Vision

Computer vision is the technology that teaches machines how to understand what they are looking at. You need to consider how human vision works, it is clear to the eyes that this is a worker wearing a protective helmet. Computer vision does the same thing, it changes the things it detects into data that the system can understand [19].

The process is straightforward:

- Vision: The camera is the eyes that captures objects on the construction site.
- Understanding: This refers to the machine's ability to identify objects in an image, such as a worker, safety equipment, or a missing piece of safety clothing.
- Action: The system use this understanding to perform some tasks, like safety PPE violations or sending an alert about an incident involving a worker.

The Computer Vision Pipeline

The new computer vision system is truly amazing; it's not just a simple matter of looking at an image, it's so intelligent that it processes it through a whole series of steps. The steps for a system detecting safety equipment is usually like this [19]:

- **Data Collection & Annotation:** Lots of photos and video frames are collected from building sites. People mark the boxes around things like hard hats and vests to create a useful dataset.
- **Preprocessing & Augmentation:** Data is standardized by reducing image size and adjusting lighting to enhance the accuracy of the model. Images are also artificially enhanced by rotating them, changing their colors, or adding artificial lighting to simulate different times of day and various weather conditions.
- **Feature Extraction & Inference:** Images are fed into a neural network called (such as YOLO or CNN) a filter, where the network layers act as filters. These filters extract simple values, such as object edges, in the basic upper layers, and extract complex features from objects in the deeper layers of the model.
- **Post-Processing:** The algorithms produce multiple predictions for the squares surrounding a single object. A technique known as "homogeneous maximum braking" (NMS) which in turn retains only the expected correct results and excludes the incorrect ones, in order to keep only the correct squares for each object that is detected. [19]

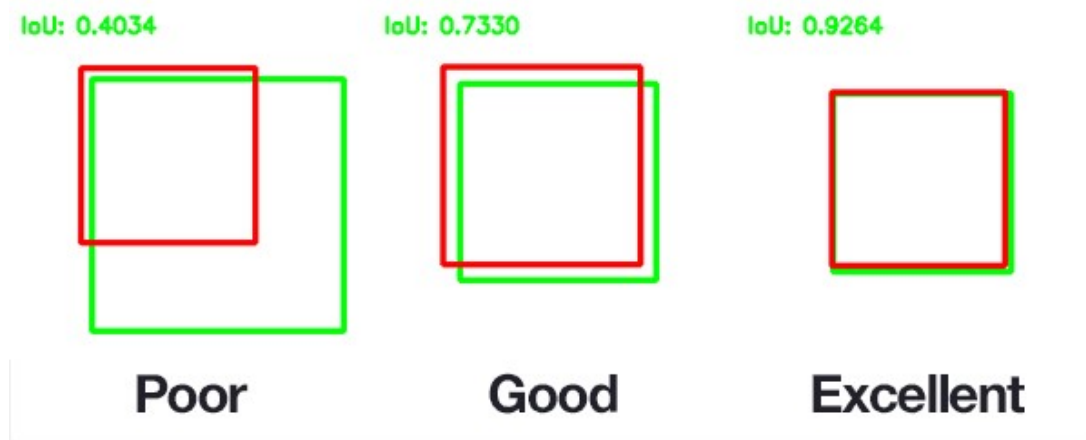


Figure 2.4: Visualizing Intersection over Union (IoU) scores.

2.2.2 Deep Learning

Deep learning can accurately and continuously identify safety violations and hazardous situations involving workers on construction sites by automatically extracting convolutional neural networks (CNN).

In AI helps recognize the physical movement patterns of workers, rather than just their standing or stationary state, it can also recognize activities such as climbing ladders, operating machinery, identifying equipment, bending, or

falling. To understand why Deep Learning (DL) is used here instead of traditional Machine Learning (ML), it's important to look at how they learn [13].

In ML, we are basically giving a computer a very specific checklist to train an ML system to recognize for example a helmet, first has to tell the computer what to look for "Look for something yellow, with curved edges, that sits on top of a person, this step is called "feature extraction" [13].

However, in the domain of DL, the process can be likened to presenting a computer with a thousand images of individuals wearing hard hats, and a thousand images of individuals not wearing helmets. There is no necessity for the employment of a checklist, the programme is capable of identifying patterns independently, including the curve, the colour, and the placement of objects. In machine learning (ML), the algorithm (1) used in a simple linear regression model makes a prediction (y) by multiplying the input data (x) by a learned weight (w) and adding a bias (b):

$$y = wx + b \quad (1)$$

DL uses this elementary equation within a series of layers known as a Convolutional Neural Network (CNN) and DL uses an activation function (like ReLU) to pass the data through a single straight line all these to deal with non-linear real-world images. The equation (2) for a single neuron in a deep network looks like this [18]:

$$f(x) = \max(0, Wx + b) \quad (2)$$

CNNs are capable of this because they have millions of these interconnected math operations, they can automatically detect the difference between a worker safely climbing a ladder and a worker dangerously not complaining.

A. YOLO (You Only Look Once)

is selected as the primary object detection algorithm because it performs object localization and classification simultaneously in a single forward pass. This enables fast real-time detection of workers, helmets, safety vests, electric drills, hand saws, heavy machinery, and other construction-related objects while maintaining high detection accuracy [9.10.11].

It is a fact that older algorithms used to scan an image multiple times in different chunks to find objects, it was accurate but very slow. YOLO changed the meaning in AI by doing one thing better than anyone else, looking at the whole image in one go, it divides the image into a grid and predicts where objects are (bounding boxes) and what they are (classifications) in a single mathematical sweep.

When YOLO looks at a grid square, it calculates a Confidence Score for any object it finds. This score answers two questions: I'm so excited to find out how sure i am that an object is here, and i can't wait to see how accurate my bounding box is. It calculates (3) this using Intersection over Union (IoU), which measures how perfectly the AI's predicted box overlaps with the actual object (4):

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (3)$$

The final confidence score equation is:

$$\text{Confidence} = \text{Pr}(\text{Object}) \times \text{IoU}_{\text{pred}}^{\text{truth}} \quad (4)$$

YOLO versions

- YOLOv1 to v3: The foundation is now set. AI is incredible at detecting objects in real-time, but it does have its limits when it comes to very small items.
- YOLOv4 to v7: incredible improvements in accuracy and speed, and its uses for industrial use.
- YOLOv8 and YOLOv10/v11 are the absolute best there is right now, they're incredibly lightweight, which means you can use them straight away on a construction site's local camera systems no massive, expensive supercomputer needed. [8]

B. Human Pose Estimation

Human Pose Estimation identifies key points on the body, such as the shoulders, elbows, hips, knees, and ankles, to estimate a worker's posture.

These skeletal landmarks help analyze worker movements, recognize abnormal body positions, and support fall or injury detection [9] [10].

The AI is aware of the location of the joints, it then connects the dots to build a stick figure, the system uses straightforward equations to analyze behavior.

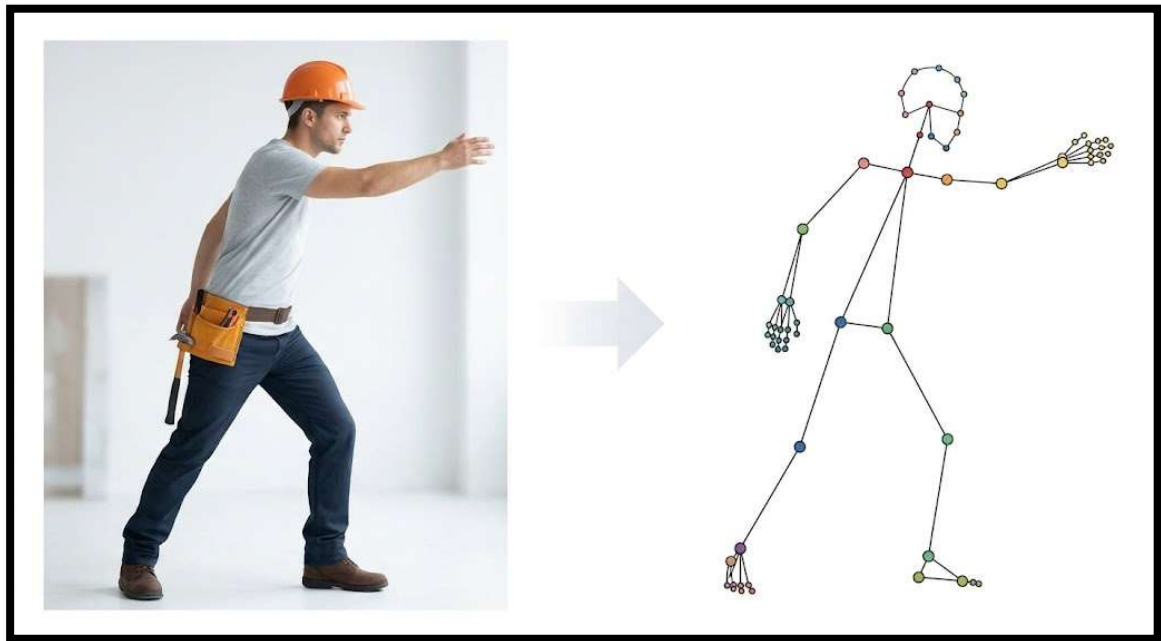


Figure 2.5: Schematic overview of a deep learning flow for construction safety monitoring

- **Ergonomic danger:** If the hip, knee and ankle joints are positioned too close together and for too long, the system will alert the worker to the fact that they are lifting with their back instead of their legs [11].
- **Fall Detection:** if the Y (height) values for the head and shoulder joints suddenly become the same as the ankle joints, the system will immediately trigger a fall alert [12].

It's also worth mentioning that HPE models usually don't provide exact coordinates right away. But don't worry, they've got you covered. Instead, they generate a heatmap for each joint. A heatmap is a grid of probabilities showing how likely it is that a specific pixel contains a joint (such as an elbow). we know how tricky it can be to train an AI to generate accurate heatmaps, but rest assured that we use a loss function such as Mean Squared Error (MSE) to help with that. We've worked out this equation (5) to make sure that the AI is given a little penalty when its predicted joint location ($\hat{y}_i$) is found to be a significant distance away from the actual true joint location (y_i):

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (5)$$

The wonderful thing is, by constantly reducing this error as you train, the AI can learn to accurately see what a worker's posture is in real-time, even when they are wearing bulky safety gear or carrying heavy materials.

C. Dataset Integration and labeling

Depending on a multi-datasets approach will allow all the objectives of this research to be achieved. This will include a dataset of images of various tools, tool detection, sharp tools, drills and hand saws, and a dataset of large construction vehicles. Combining the different datasets will help to increase the system's ability to recognize different types at the same time. Although the labelling of datasets can appear to be a tiresome task, it is in fact the most significant stage within the project. If the AI is exposed to suboptimal examples, it will be prone to erroneous predictions, even if the underlying maths model is robust [4]. It is recommended that visual annotation tools (VAT) such as the Computer Vision Annotation Tool (CVAT) or Roboflow be used in place of manual coordinate entry. The user simply uploads images and points at the mouse to draw boxes or place dots, and the software automatically generates the files required by the neural networks [5].

The following is the precise structure required for those files when used on construction sites:

- **Labeling for YOLO**

When teaching YOLO to find helmets, drills and heavy machinery, we are drawing boxes around the items to show where they are. For every image, YOLO requires a text file with the same name (e.g., worker_01.txt). Every object that are labelled in the text file will have one line of numbers inside it that looks like this:

[Class_ID], [Center_X], [Center_Y], [Width] and [Height].

Class_ID: A simple number representing the category (e.g., 0 for Worker, 1 for Helmet, 2 for Drill).

The Maths (normalisation): YOLO does not use raw pixels; instead, it converts your box into a decimal value between 0.0 and 1.0, depending on the size of the image. If a drill is in the exact centre of the photo, its Center_X and

Center_Y values will be 0.5, this enables the boxes to automatically adjust their size to ensure they remain perfectly wrapped around the objects, even if the AI later resizes the camera feed to speed up processing [12].

- **Labelling for Human Pose Estimation**

Pose estimation is a lot more detailed than just drawing a box. Basically, we are connecting the dots across the human body to map out a worker's posture. Most pose estimation models use the COCO format to store all the labels for your entire dataset in a single (.json) file. This is better than storing them in individual text files. For each person, the system records an array of keypoints [11]. These are typically 17 standard joints Figure 2.5:

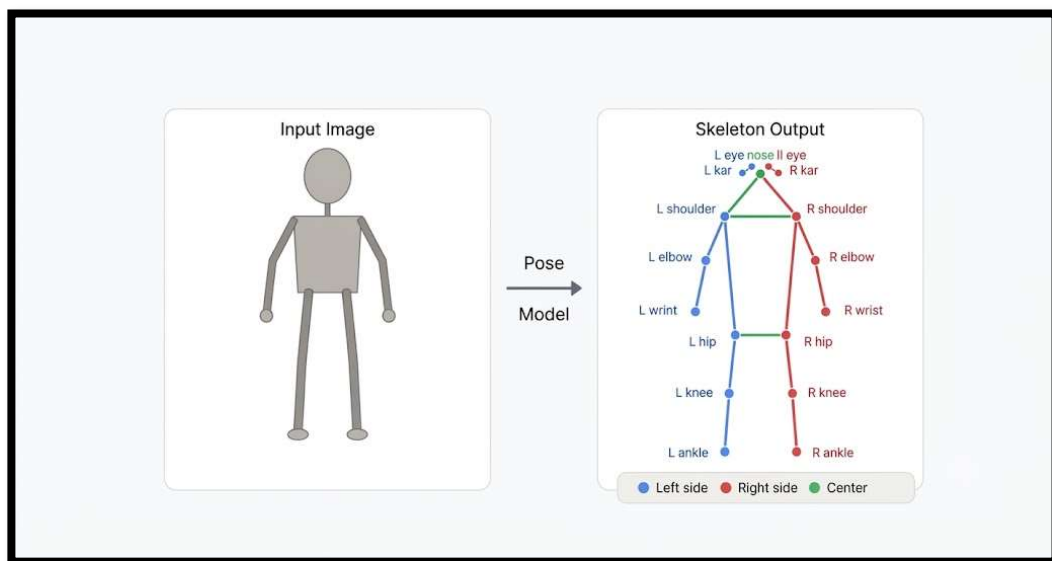


Figure 2.6: Mapping human keypoints for pose estimation

Each joint is defined by three values: $[x, y, v]$:

x and y : means need the exact pixel locations of that joint.

V is for Visibility: is the main weapon for construction sites, where workers are constantly moving behind machinery or materials and there are there results:

- $V = 0$: The joint is not visible in the photo (e.g., the camera cuts off at the worker's waist).
- $V = 1$: The joint is in the frame but hidden (e.g. the worker's leg is behind a concrete block). The AI will use the rest of the body to mathematically guess where the leg is.
- $V = 2$ It is clear to the camera that the joint is visible.

- **The Multi-Dataset:**

By depending on a multi dataset, all the objectives of this research can be achieved, which includes a dataset of images of various tools, tool detection, sharp tools, drills, hand saws, person fallen and person walking normally, and a dataset of large construction vehicles, as combining the different datasets helps to increase the system's ability to recognize different types at the same time. [1][3][5][12][13].

In recent years, computer vision researchers have completed the development of models and their transition from primitive to advanced stages that serve the field of safety in construction in particular, such as monitoring the movement of workers and the types of safety equipment used, which in turn aim to raise the level of protection of workers in large and complex construction areas and also reach huge industrial cities.

2.3 Related Work

During the writing of this research, safety at construction sites was a major focus for researchers, particularly regarding the behavior of construction workers with hazardous equipment and their disregard for safety regulations. Several machine learning models, including computer vision and deep learning, were presented at the time for automating safety measures and improving the prevention of construction accidents.

And to clarify and differentiate the related work, they will be categorized into two parts:

2.3.1 Safety Monitoring and PPE Detection

This section will specialize in studies related to compliance with the use of safety and PPE at construction sites, detection of hazardous tools, and unsafe movements between workers and machines:

Rahman et al. (2026) Techniques YOLO11 (YOLO11x), SMOTE balancing, and data augmentation, dataset CHVG and Roboflow, Objective to create an automated

real-time Personal Protective Equipment (PPE) compliance detection system, Results reached a mAP50 of 96.9% and an inference time of 7.3 ms, Limitation requires high end hardware for low inference time, and only detects main PPE classes, missing smaller tools. [1]

Ghatfan Emery Razan (2026) Techniques YOLOv11 based DL models, dataset SH17 Dataset heavily augmented, Objective To present an intelligent automated object-detection system utilizing the new YOLOv11 architecture to monitor specific PPE, Result system achieved a total precision and mAP@0.5 of 81.0%, showing strongest performance on masks and hardhats, Limitation model experienced decreased confidence when detecting smaller, ambiguous objects and suffered from susceptibility to visual noise, such as motion blur or grayscale conditions. [2]

Burcu Çarklı Yavuz. (2026) Techniques YOLOv11 and YOLO26 architectures evaluated across 30 different configurations, dataset SH17 alongside CHV and SHEL5K, Objective To provide evidence-based deployment guidelines for real-time safety monitoring on edge devices by benchmarking model scales, Results YOLOv11 is explicitly recommended for latency-constrained edge scenarios when using the "nano" or "small" scale models, Limitation It must be treated as a preliminary finding that requires much broader testing to be proven true. [3]

S. Malaikrisanachalee (2026) Techniques YOLO architecture integrated with an ESPCN super resolution module, dataset SH17 and custom low light imagery, Objective to solve the challenge of detecting very small PPE items captured by long distance cameras in poor lighting environments where standard YOLO models lose confidence, Results Achieved an mAP@0.5 of 92.2% on enhanced images, Limitation Requires the additional computational overhead of running a super resolution module before the YOLO network processes the image, which can create a bottleneck for real-time edge processing. [4]

Suman Raj and Bhavani A. Madhabhavi (2025) Techniques YOLOv11-m evaluated on edge accelerators and high-end GPU workstations, Dataset SH17, Objective to benchmark the inference times and accuracy latency trade-offs of Deep Neural Networks designed to detect unique visual identifiers, Results achieved 93% precision on a YOLOv11-m retrained on the smaller datasets, After some more improvements, the precision went up to 99.5% DNN models, Limitation find it hard

to identify people in busy or changing outdoor places this is because there are not many sets of data that include different types of lighting and weather. [5]

Khoiril Umat, Gunawan (2025) Techniques YOLOv10 with multi version, Dataset SH17 (targeted only 1620 Image), Objective to evaluate the YOLOv10 architecture's ability to act as a lightweight, real-time object detection system, Results YOLOv10-M model achieved a peak mAP50 of 98.4% when using an 80:10:10 data split, Limitation limiting the data scale constrains the model's ability to generalize across highly complex, cluttered construction sites where workers are heavily occluded or overlapping. [6]

Surya N. Reddy (2024) Techniques: Activity Recognition Models combined with Object Detection Techniques, Dataset SH17, Objective to reduce false alarms in PPE monitoring by developing a system that first recognizes the specific action a worker is performing, Results this led to a 23% improvement in the F1-score compared to the standard, Limitation it remains difficult to understand human actions within a complex shop floor environment. [7]

Hafiz Mughees Ahmad (2024) Techniques YOLOv9-e, dataset SH17, Objective to introduce the SH17 dataset and establish a deep learning baseline for non-invasive, automated PPE compliance, Results achieved an overall mean average precision (mAP) of 70.9%, Limitation Accuracy dropped significantly for minority classes as items with fewer image samples, like face guards. [8]

2.3.2 Injury and Fall Detection & Prediction

This section will specialize in studies related to determining human body posture, analyzing abnormal postures, detecting and predicting falls, and addressing potential injuries at construction sites:

K. Jishnuraj, M. Vergin Raja Sarobin, et al. (2026), Techniques FallCNN architectures paired with transfer learning models, Dataset UR Fall Detection, Objective to build highly accurate fall detection by transferring deep learning weights from established convolutional networks, The results of using FallCNN structures together with transfer learning models had an accuracy of 97.66%, Limitation of combining dense convolutional structures with transfer learning creates a burden on

the device's computing power, which makes it difficult to process things in real time on devices that don't have a lot of power [9].

Suraiya Akter, Nahid Nahar. (2025) Spatiotemporal Context Tracking, techniques Vision Transformers (ViT) combined with (LSTM) networks, Dataset UR Fall Detection, Objective to monitor worker movements in the work environment and the worker's speed of movement, which allows us to predict before a worker falls, Results by using the Vision Transformer and the LSTM to the model achieved an impressive 98.45% accuracy with an Area Under Curve (AUC) of 0.992, Limitation The framework found it hard to separate actual falls from quick bending movements, which triggered false alarms [10].

Koffi , (2025) Improving Dataset Quality for Impact Detection, Techniques 3D Skeleton Extraction (Mediapipe BlazePose) integrate with Machine Learning (Random Forest, SVM) and Deep Learning (CNN, LSTM), Dataset UP-Fall Detection, Objective clean and restructure the dataset by extracting precise 3D skeletal joints and relabelling the exact frames where an impact occurs, Results after cleaning the data and focusing on the skeletal coordinates beter than raw video pixels the researchers achieved over 98% accuracy across multiple deep learning and machine learning algorithms, Limitation on this research only enhanced a small subset of the original dataset (processing just 5 subjects performing 5 scenarios) [11].

Cho & Solak, (2025) Anticipatory Pre-Fall Prediction via Pose Forecasting, Techniques hybrid dynamic graph neural networks (DGNN) and (LSTM), Dataset UR Fall Detection, (URFD) and Berkeley MHAD, Objective to recognizing and analyse the brief "transient state" between stability and an impending fall by forecasting exactly where the worker's skeletal joints would be in future frames, Results the hybrid architecture achieved an exact fall detection accuracy of **98.2%**, limitation If the CCTV video feed lags, drops frames, or stutters the LSTM temporal sequence breaks down, causing the future pose predictions to fail entirely [12].

Kong, V. (2025) Techniques YOLOv11 with SVM, dataset UR Fall Detection & MCFD, Objective to extract nine critical geometric features (such as center of mass and body angles) to classify a physical collapse, Results Achieved 88.8% precision, 94.1% recall, an F1-score of 91.4%, and a specificity of 95.6%, Limitation the system requires aggressive validation methods to analyze sequential frame changes to ensure robust detection [13].

Hui Ren, Ping Lan. (2025), Techniques BMR-YOLO, dataset UR Fall Detection, Objective to achieve effective fall detection in complex, obstructed environments while reducing computational load, Results achieved an accuracy of 95.24% on the UR Fall external validation sequences, Limitation severe falls that are completely occluded by crowds still result in missed detections [14].

Ekram Alam, Abu Sufian. (2024) Steady-Unsteady-Fallen Posture Model, Techniques Video classification and low-resolution data processing without compression, dataset GMDCSA-24, Objective to provide diverse indoor scenarios addressing false positives triggered, Results generated a real-time capable dataset using low-cost 0.92 Megapixel webcams, increasing system robustness against false positives, Limitation the dataset was generated using simulated falls performed by only four actors in three specific home setups, which may present generalization challenges [15].

Guo et al. (2018) A proposed skeleton-based, real-time identification method combines computer vision, safety knowledge and ergonomic theory. It identifies unsafe behaviours by simplifying high-dimensional, dynamic motions into static "leading postures" that are described by parameters. This is different from traditional video tracking methods, which need a lot of data processing and result in delayed recognition. This approach doesn't get in the way and focuses only on the important joint coordinates to capture basic spatial features [16].

Liu et al. (2021) This review looked at how computer vision methods are used to monitor the unsafe actions of construction site workers. It also assessed the status of the methods, historical explorations and ongoing technical challenges. Whilst other studies have focused on a limited number of areas, this research has drawn upon all the extant information. The study examined the behaviour of workers in the following areas: the wearing of the correct protective equipment, posture, and interaction with the immediate environment. The paper confidently identifies key areas for research and future development in this field. A tool called VOS Viewer is used to analyse 122 important academic articles from 2000 to 2021, thus enabling the identification of relevant sources and the establishment of a solid foundation for further research [17].

Author	Year	Techniques	Dataset	Objective	Results	Limitation
[1] Rahman et al.	2026	YOLO11 (YOLO11x), SMOTE balancing, and data augmentation	CHVG and Roboflow dataset	To create an automated real-time Personal Protective Equipment (PPE) compliance detection system	Reached a mAP50 of 96.9% and an inference time of 7.3 ms	Requires high end hardware for low inference time, and only detects main PPE classes, missing smaller tools
[2] Ghatfan Emery Raza	2026	YOLOv11 based DL models	SH17	monitor specific PPE adherence in real-time within the construction sector.	achieved a total precision and mAP@0.5 of 81.0%	Experienced decreased confidence when detecting smaller, ambiguous objects
[3] Burcu Çarklı Yavuz	2026	YOLOv11 and YOLO26 architectures	SH17, CHV and SHEL5K	provide evidence-based deployment guidelines for real-time for PPE	YOLOv11 is explicitly recommended for latency constrained edge scenarios when using the nano or small scale models	It must be treated as a preliminary finding that requires much broader testing to be proven true
[4] S. Malaikrisa nachalee	2026	YOLO integrated with an ESPCN	SH17	to solve the challenge of detecting very small PPE items	Achieved an mAP@0.5 of 92.2% on enhanced images	Requires the additional computational overhead of running a super resolution module before the YOLO network processes the image
[5] Suman Raj, Bhavani A. Madhabhavi	2025	YOLOv11-m	SH17 and Ocularone Dataset	To benchmark the inference times and accuracy-latency	Achieved 93% precision on a YOLOv11-m	Standard DNN models struggle to uniquely identify subjects in crowded environment
[6] Khoiril Umat, Gunawan	2025	YOLOv10 with multi version	SH17	to evaluate the YOLOv10 architecture's ability to act as a lightweight, real-time object detection system	YOLOv10-M model achieved a peak mAP50 of 98.4%	limiting the data scale constrains the model's ability to generalize across highly complex, cluttered construction sites where workers are heavily occluded or overlapping.
[7] Surya N. Reddy, Vaibhav Kurrey	2024	Activity Recognition Models integrated with Object Detection	SH17	To reduce false alarms in PPE monitoring	Led to a 23% improvement in the F1-score compared to standard	Understanding human actions within a complex shop floor environment remains difficult
[8] Hafiz Mughees Ahmad and	2024	YOLOv9-e	SH17	To introduce the SH17 dataset and establish a deep learning baseline	mean average precision (mAP) of 70.9%	Accuracy dropped significantly for minority classes

Afshin Rahimi						
[9] K. Jishnuraj, M. Vergin Raja Sarobin, et al.	2026	FallCNN architectures paired with transfer learning models.	UR Fall Detection	to build highly accurate fall detection by transferring deep learning weights from established convolutional networks	achieved a 97.66% accuracy	combining dense convolutional architectures with transfer learning creates computational drag, making real-time processing on low-power devices difficult
[10] Suraiya Akter, Nahid Nahar, et al.	2025	Vision Transformers (ViT) and LSTM	UR Fall Detection	To develop a lightweight, real-time fall detection	achieved 98.45% accuracy on fall detection while minimizing false detection rates.	highly dependent on specific camera viewpoints only
[11] Koffi	2025	3D Skeleton Extraction, integrate with ML (Random Forest, SVM) and DL (CNN, LSTM)	UP-Fall Detection	clean and restructure the dataset by extracting precise 3D skeletal joints and relabelling the exact frames where an impact occurs	after cleaning the data and focusing on the skeletal coordinates better than raw video pixels, the researchers achieved over 98%	only enhanced a small subset of the original dataset
[12] Cho & Solak	2025	hybrid dynamic graph neural networks (DGNN) and (LSTM)	UR Fall Detection and Berkeley MHAD	to recognizing and analyse the brief transient state between stability and an impending fall	hybrid architecture achieved an exact fall detection accuracy of 98.2%	If the CCTV video feed lags, drops frames, causing the future pose predictions to fail entirely
[13] Kong, V., et al	2025	YOLOv11 with SVM	UR Fall Detection & MCFD	to extract nine critical geometric features (such as center of mass and body angles) to classify a physical collapse.	achieved accuracy 88.8%	the system requires temporal validation methods to analyze sequential frame changes to ensure robust detection.
[14] Hui Ren, Ping Lan	2025	BMR-YOLO	UR Fall Detection	to achieve effective fall detection in complex, obstructed environments while reducing computational load	achieved an accuracy of 95.24% on the UR Fall external validation sequences.	severe falls that are completely occluded by crowds still result in missed detections.
[15] Ekram Alam, Abu	2024	Video classification and low-resolution data processing	GMDCSA-24	to provide diverse indoor scenarios	generated a real-time capable dataset using low-cost 0.92	the dataset was generated using simulated falls

Sufian, et al.		without compression.		addressing false positives triggered	Megapixel webcams, increasing system robustness against false positives.	performed by only four actors in three specific home setups, which may present generalization challenges.
[16] Liu et al.	2021	Bibliometric and content based literature review	122 relevant academic papers published between 2000 and 2021	To review the historical explorations, status, and challenges of using computer vision for worker safety	Mapped research hotspots such as object detection, object tracking, and action recognition	Revealed persistent industry wide limitations, including visual occlusion, small object detection errors, and irregular tracking
[17] Guo et al. 1	2018	Depth-image skeletonization, parameterization of leading postures, and Kinect 2.0	Chinese Accident Case Collection templates and a custom laboratory test with 9 testes	To achieve real time identification of unsafe behaviors by simplifying dynamic motions into static postures	Successfully simplified complex posture tracking down to 6 key joint parameters for instant identification	Limited sample size, hardware interference from outdoor sunlight, and lack of testing on multiple overlapping workers

Table 2.1: Summary of Related Work on Construction Site Safety Monitoring

References

- [1] Rahman, A., Ahmed, M. S., AlBugami, K. N., Alabbad, A. Y., AlFantoukh, A. A., Alshaikhahmed, Y. H., Alzahrani, Z. S., Khan, M. A. A., Youldash, M., & Alshahrani, S. M. (2026). PPE-EYE: A Deep Learning Approach to Personal Protective Equipment Compliance Detection. *Computers*, 15(1). DOI: <https://doi.org/10.3390/computers15010045>
- [2] G. E. Razan, D. Kurnianingtyas, M. H. Shodiq, S. F. M. R. P. Sitompul, A. S. Lopulalan, and K. Arai, "Utilizing YOLOv11 for Real-Time Construction PPE Compliance Detection", *Eng. Technol. Appl. Sci. Res.*, vol. 16, no. 2, pp. 33866–33873, Apr. 2026. DOI: <https://doi.org/10.48084/etasr.15801>
- [3] Çarklı Yavuz, B. Scale-Dependent Performance Analysis of YOLO26 and YOLOv11 for PPE Detection. *Electronics* 2026, 15, 1146.. DOI: <https://doi.org/10.3390/electronics15061146>
- [4] Malaikrisanachalee, S.; Wongwai, N.; Kowcharoen, E. ESPCN-YOLO: A High-Accuracy Framework for Personal Protective Equipment Detection Under Low-Light and Small Object Conditions. *Buildings* 2025, 15, 1609. DOI: <https://doi.org/10.3390/buildings15101609>
- [5] Raj, S., et al. (2025). Ocularone-Bench: Benchmarking DNN Models on GPUs to Assist the Visually Impaired. *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2504.03709>
- [6] Gunawan, W., & Khoiril Umat, Y. (2025). YOLOv10 for Real-Time Detection of Personal Protective Equipment on Construction Workers. *ILKOM Jurnal Ilmiah*, 17(2), 131-139. DOI: <https://doi.org/10.33096/ilkom.v17i2.2544.131-139>
- [7] Reddy, S. N., Kurrey, V., Nagar, M., & Gupta, G. R. (2024). Action Recognition based Industrial Safety Violation Detection. *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2412.05531>
- [8] Ahmad, H. M., & Rahimi, A. (2024). SH17: A dataset for human safety and personal protective equipment detection in manufacturing industry. *Journal of Safety Science and Resilience*. DOI: <https://doi.org/10.1016/j.jnlssr.2024.09.002>
- [9] Jishnuraj, K., Sarobin, M. V. R., Anbarasi, J., Jasmine, G., & Panjanathan, R. (2026). Fall detection among elderly persons using FallCNN and transfer learning models. *Frontiers in Artificial Intelligence*, 9, 1734096. DOI: <https://doi.org/10.3389/frai.2026.1734096>
- [10] Akter, S., Nahar, N., Imam, M. H., & Khandaker, M. U. (2025). A Real-Time Fall Detection Framework Using Vision Transformer and LSTM for Elderly People. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM)*, 37, 903-913. DOI: <https://doi.org/10.55549/epstem.1358>
- [11] Koffi, T. Y., Mouchid, Y., Hindawi, M., & Dupuis, Y. (2025). An improved 3D skeletons UP-Fall dataset: enhancing data quality for efficient impact fall detection. *Proceedings of SPIE*

13517, Seventeenth International Conference on Machine Vision (ICMV 2024), 135170R. **DOI:** <https://doi.org/10.1117/12.3055191>

[12] Cho, Y., & Solak, G. (2025). Anticipatory Fall Detection in Humans with Hybrid Directed Graph Neural Networks and Long Short-Term Memory. arXiv preprint. **DOI:** <https://doi.org/10.48550/arXiv.2502.19048>

[13] Kong, V., et al. (2025). PIFR: A novel approach for analyzing pose angle-based human activity to automate fall detection in videos. PLOS One. **DOI:** <https://doi.org/10.1371/journal.pone.0325253>

[14] Ren, H., & Lan, P. (2025). BMR-YOLO: A deep learning approach for fall detection in complex environments. PLoSOne. DOI: <https://doi.org/10.1371/journal.pone.0335992>

[15] Alam, E., Sufian, A., Dutta, P., Leo, M., & Hameed, I. A. (2024). GMDCSA-24: A dataset for human fall detection in videos. Data in Brief, 57, 110892. **DOI:** <https://doi.org/10.1016/j.dib.2024.110892>

[16] Liu, W., Meng, Q., Li, Z., & Hu, X. (2021). Applications of Computer Vision in Monitoring the Unsafe Behavior of Construction Workers: Current Status and Challenges. Buildings, 11(9), 409. **DOI:** <https://doi.org/10.3390/buildings11090409>

[17] Guo, H., Yu, Y., Ding, Q., & Skitmore, M. (2018). Image-and-skeleton-based parameterized approach to real-time identification of construction workers' unsafe behaviors. Journal of Construction Engineering and Management, 144(6), 04018042 **DOI:** <https://doi.org/10.3390/computers15010045>

[18] Rothstein, M. A. (2020). The Occupational Safety and Health Act at 50: Introduction to the Special Section. American Journal of Public Health, 110, 613-614. **DOI:** <https://doi.org/10.2105/ajph.2020.305623>

[19] Neha, F., et al. (2024). From classical techniques to convolution-based models: A review of object detection algorithms. arXiv preprint. **DOI:** <https://doi.org/10.48550/arXiv.2412.05252>