Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING ENABLED VISION TRANSFORMER FRAMEWORK FOR WORKER SAFETY MONITORING IN CONSTRUCTION SITES

Dr. P. PAVAN KUMAR[1], RAKSHITHA KIRAN P[2], DANAM NOELLE[3], V. V. RAMA KRISHNA[4], C. RAGHAVENDRA[5], GANGULA SURENDAR[6]*, KOMALI GOVINDU[7], NARESH ALAPATI[8]

[1]Department of Computer Science and Engineering (AI & ML), Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana - 501506, India
[2]Department of Computer Science and Engineering (Cyber Security), Dayananda Sagar College of Engineering, Bangalore, Karnataka 560011, India
[3]Department of Civil Engineering, Aditya University, Surampalem, Andhra Pradesh, India
[4]Department of Electronics and Communication Engineering, Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India
[5]Department of Computer Science and Engineering (Cyber Security), CVR College of Engineering, Hyderabad, Telangana, India
*[6]Department of Artificial Intelligence and Data Science, GMRIT Deemed to be University, Rajam, Andhra Pradesh 532127, India
[7]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522502, India
[8]Department of Computer Science and Engineering, Vignan's Nirula Institute of Technology and Science for Women, Pedapalakaluru, Andhra Pradesh, India

E-mail: [1]pavan.aidsgnitc@gniindia.org, [2]rakshitha610@gmail.com, [3]noelle196q1d1905@gmail.com, [4]vvrk25@gmail.com, [5]crg.svch@gmail.com, *[6]surendar.g@gmritdu.edu.in, [7]komaligovindu1996@gmail.com, [8]alapatinaresh13@gmail.com

ABSTRACT

Construction sites are dangerous places where employees are often at risk in various ways, including falling objects, unsafe equipment interactions and failure to wear personal protective equipment (PPE) as specified in the regulations. Manual safety monitoring is typically ineffective, unreliable and not able to conduct hazard analysis in real-time. This paper aimed to overcome these drawbacks by introducing a Deep Learning Enabled Vision Transformer Framework to monitor worker safety in construction sites. The suggested solution combined CNN with VC with Adaptive Contextual Risk Assessment (ACRA) to detect PPE compliance, unsafe activities, restricted area intrusion, and hazardous equipment proximity in real time. A hybrid construction safety database with 48,732 annotated images was used to evaluate the proposed approach under various scenarios, such as occlusion of workers, illumination changes, and complex scenes. The results of the experiments have shown the proposed framework to be superior to the CNN, ResNet50, Faster R-CNN, YOLOv5, EfficientNet and Swin Transformer models in terms of accuracy (98.1%), precision (97.3%), recall (96.8%), F1 score (97.0%), and mean Average Precision (mAP) (97.6%). Under complex construction environments, the Vision Transformer (ViT) encoder significantly enhanced its understanding of context hazards and reduced false alarm proportions. The proposed framework was also capable of achieving real-time processing capability with a low detection latency that is suitable for industrial deployment. Incorporating Vision Transformers and adaptive contextual risk assessment can significantly enhance intelligent construction worker safety monitoring and facilitate automated industrial surveillance for accident prevention and workplace safety improvement, the study showed.

Keywords: *Construction Safety Monitoring, Vision Transformer, Deep Learning, PPE Detection, Hazard Detection, Worker Safety Surveillance*

1. INTRODUCTION

The construction sector is an important component of the development of the world economy and the expansion of the infrastructure. Although the days of traditional project management and construction automation have been superseded, construction sites are still one of the more dangerous working environments, with the presence of heavy machinery, elevated structures and moving equipment, as well as unsafe working conditions [1]. The statistics of occupational safety report that in the world construction work-related accidents make a significant percentage of the total of workplace injuries and fatalities [2]. The main reasons for accidents are unsafe working practices, inadequate supervision "on the spot", improper use of personal protective equipment (PPE) and late detection of hazardous situations [3]. The problems highlight the need for smart and automatic systems for worker safety monitoring that can enhance the safety of construction sites and prevent accidents.

The traditional construction safety management is mainly based on manual inspection and human supervision [4]. Safety officers check workers around the clock for compliance with safety rules that include safety vest and helmet wearing, and adherence to restricted areas. Manual monitoring systems have a number of drawbacks, however, such as human fatigue, the variability of observations made, the delay in responding and the lack of capacity to monitor large-scale construction sites constantly [5]. Besides, conventional monitoring systems rely heavily on the expertise of humans, and also cannot offer accurate real-time hazard analysis under dynamic construction conditions [6]. Hence, there is a growing need for intelligent safety monitoring systems that can provide automated monitoring and rapid hazard detection.

In recent years, the field of Industrial monitoring and Computer vision has been revolutionized by the recent developments of Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) [7]. The deep learning methods showed outstanding results in image classification, object detection and activity recognition. Of these, Convolutional Neural Networks (CNNs) were far more popular as they could automatically learn hierarchical visual features from images [8]. CNN-based architectures, including AlexNet, VGGNet and ResNet were used in a few studies to detect construction workers [9] and monitor PPE compliance and unsafe activities. These methods were found to be more accurate than the conventional handcrafted feature extraction methods in the context of safety monitoring.

Additionally, object detection systems like Faster R-CNN, SSD, and YOLO greatly improved the efficiency of worker safety monitoring in the real world [10]. Similar work in detecting workers, helmets, safety vests and hazardous activities has been carried out in surveillance videos with YOLO-based models, which were found to be both faster in inference and more accurate [11]. CNN-based and object detection methods are still plagued by various issues in real construction environments with complex background, occlusion of workers and illumination difference, and dense construction areas [12]. Traditional convolution operations tend to focus on local feature extraction and are not able to learn long-range contextual dependencies that are needed for a full understanding of a scene [13].

Recently, transformer-based architectures have gained significant popularity in computer vision applications as a powerful alternative to the traditional CNN-based architectures, tackling these shortcomings. To address these challenges, transformer-based architectures have recently become popular alternatives in computer vision tasks. The Vision Transformers (ViTs) are models inspired by the transformer architectures used in NLP research, where they models global contextual relationships among words [14]. The Vision Transformers are good at capturing long-range dependencies and contextual interactions in complex scenes, unlike traditional CNN architectures. This is particularly useful for construction safety monitoring systems, where the context of relationships among workers, machines, and environments are crucial for hazard identification [15].

The transformer has been shown to achieve superior results in various image recognition, anomaly detection, and industrial inspection tasks in recent studies [16]. Swin Transformer and hybrid transformer-based models were found to be more robust models under different illumination conditions and partial occlusion of the objects [17]. Moreover, feature extraction approaches based on a transformer have improved the contextual awareness and detection stability in cluttered industrial settings [18]. But, the use of Vision Transformers for smart worker safety monitoring in construction sites is still relatively little explored. Most systems existing in the market only detect the presence of PPE in an isolated manner without conducting a contextual hazard

analysis which also analyses worker posture, the intrusion into restricted zones and the proximity to hazardous equipment [19].

One of other greatest problems with the current construction safety monitoring systems is that they experience excessive false alarms when operating in dynamic environment [20]. Accurate safety monitoring is hard to achieve in the construction industry due to the presence of dust, smoke, overlapping work activities, poor visibility and changing weather conditions [21]. From there, current CNN systems often suffer from poor generalization ability and poor detection performance. It is therefore an urgent requirement to provide an intelligent monitoring framework that can adaptively learn context, extract multi-scale features and automatically assess risk in real time.

Despite significant advances in computer vision, deep learning, and construction safety monitoring systems, achieving accurate real-time worker safety monitoring in dynamic construction environments remains a major challenge. Existing approaches often suffer from limited contextual understanding, poor detection performance under occlusions and varying environmental conditions, excessive false alarms, and inadequate integration of risk assessment mechanisms. These limitations reduce monitoring reliability and may compromise worker safety. Therefore, there is a need for an intelligent framework capable of adaptive contextual learning, robust multi-scale feature extraction, and real-time safety event detection for effective construction site monitoring.

In this regard, this paper introduces the Deep Learning Enabled Vision Transformer Framework for Worker Safety Monitoring in Construction Sites to tackle the above research gaps. The proposed mechanism combines the adaptive multi-label safety event detection with the Vision Transformer (ViT) based contextual feature learning to achieve intelligent real-time surveillance. The framework aimed at pinpointing the following types of PPE violations, unsafe worker behaviors, restricted zone intrusions, and hazardous proximity conditions in surveillance video streams. The proposed model differs from traditional CNN-based models by introducing self-attention mechanism, which helps to better capture the context and boosts the robustness of the detection process in challenging construction environments.

This research has several main aims, which are summarized below:

1. To establish an intelligent monitoring system for workers' safety in construction sites through the Vision Transformer.
2. To enhance the accuracy of PPE detection in occlusion, and illumination variation and crowded scene conditions.
3. To conduct a context hazard analysis to identify unsafe worker behaviours and hazardous proximity events.
4. To lower false alarms with adaptive attention-based feature learning mechanisms.
5. To benchmark the proposed model with the existing deep learning or object detection models by using the standard evaluation metrics.

The novelty of the proposed work is that it combines a Vision Transformer architecture with an adaptive contextual risk assessment in multi-event construction safety monitoring. The proposed framework not only identifies violations of PPE but also conducts a contextual scene interpretation that interprets the scene to analyze the unsafe interaction among workers and the environmental hazards at the same time. Further, adaptive attention learning mechanisms enhance the generalization ability, detection stability and stability of the model under various construction site environments.

The rest of this paper is organized as follows. The related works about deep learning and construction safety monitoring systems based on transformer are discussed in Section 2. The proposed methodology and the proposed Vision Transformer architecture are presented in Section 3. The experimental results and comparative performance analysis are discussed in section 4. Finally, the paper is concluded and directions for future research are outlined in Section 5.

2. RELATED WORK

Construction worker safety monitoring has gained considerable attention in recent years due to the rapid advancement of Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and computer vision technologies. In the construction industry, researchers have suggested several smart monitoring systems to enhance hazard detection, PPE compliance checking, worker activity recognition and accident prevention. Current methods can be grouped into three main categories: traditional computer vision techniques,

deep learning-based object detection systems, activity recognition systems, and intelligent monitoring systems based on transformers.

Early studies of construction safety monitoring were primarily based on image processing and handcrafted feature extraction methods. The tasks involving worker detection and helmet detection often used methods based on Histogram of Oriented Gradients (HOG), Haar cascade classifiers [22] and Support Vector Machines (SVMs) [22]. Under the controlled environment, these methods achieved moderate detection accuracy, but in the dynamic construction environment, which has illumination changes, background noise and occlusion, detection accuracy drops significantly [23]. In addition, manually engineered feature extraction techniques were both highly labor intensive and not robust enough for real-time implementation in industry.

Deep learning kicked the transformation of construction safety monitoring applications to a new level. Convolutional Neural Networks (CNNs) are known for their automatic feature extraction ability which made them more effective for image classification and object recognition [24]. The CNN architecture was used to classify the workers, helmets, gloves, safety vests, and dangerous activities in the construction site surveillance videos [25]. Fang et al. designed a worker safety detection system based on CNN which can detect the PPE compliance with higher accuracy than the traditional computer vision method [26]. In the same way, Wu et al. [27] introduced a deep CNN network for safety helmet detection under different environmental conditions, and obtained encouraging results.

In addition, object detection frameworks were integrated to further improve the automated construction safety monitoring, allowing for multi-object tracking and classification of construction workers and safety equipment. Faster R-CNN was one of the first region proposal based object detection model used for construction safety application [28]. The high localization accuracy of Faster R-CNN made it difficult to be used in real-time in an industrial environment due to its high computational cost. After that, Single Shot MultiBox Detector (SSD) and YOLO based architectures enhanced the detection rate with respect to accuracy while simultaneously boosting the speed of detection [29]. YOLOv3 and YOLOv4 achieved better results in real-time PPE detection applications for helmets and safety jackets and worker tracking [30].

Helmets are prone to injuries in construction sites and have gained popularity in the industry, so several researchers dedicated their efforts to studying helmet detection systems. Nath et al. presented a helmet detection system based on YOLO for automated safety compliance monitoring using surveillance systems [31]. The proposed model had the ability to detect in real time, but it was not robust in case of crowded scenes and parts occlusions. Similarly, Mneymneh et al. suggested a multi-camera safety monitoring system for construction sites that employs deep object detection models [32]. While the framework enhanced the precision of worker localization, it was lacking in terms of understanding worker interaction and environmental hazards.

The recognition of worker activities also gained attention as a research topic in the field of intelligent construction safety system. The aim of activity recognition methods was to detect unsafe worker behaviours like climbing without support, lifting posture and unsafe interaction with the equipment. A number of studies have employed LSTM networks and Recurrent Neural Networks (RNNs) to process temporal sequence of worker activities in video streams [33]. A hybrid CNN-LSTM model for unsafe activity detection in the construction environment was proposed by Yang et al. [34]. The system worked well in analyzing activities in time, but was found to be costly in terms of computation time and was not easily scalable for large surveillance systems.

Most recent research focused on multimodal safety monitoring systems with wearable sensors, IoT, and computer vision. Ding et al. suggested an IoT-based worker monitoring system that integrates the information from sensors and hazard detection based on images [35]. The system enhanced the Environmental monitoring capability but dependency on wearable sensors made the implementation more complex and expensive. In a similar way, proximity hazard analysis and collision prevention were enabled by introducing Bluetooth and RFID-based worker tracking systems [36]. But these were not easy to scale up to large deployments in dynamic construction environments and were often a burden on additional infrastructure.

Recently, transformer-based architectures came up as powerful alternatives to CNNs for computer vision tasks. For example, Vision Transformers (ViTs) use self-attention to model the global context of image patches [37]. Unlike CNNs, the architectures of transformers are good at capturing long-range relationships and contextual interactions in complex scenes. This feature is especially important in construction safety applications where identifying hazards relies on the context, proximity of equipment to workers, and interactions among the workers.

Swin Transformer integrated hierarchical attention mechanism and shifted window operations to enhance the efficiency of feature representation [38]. Some works have been reported which compare transformer-based architectures to conventional CNNs for industrial inspection and anomaly detection tasks, and showed that transformer-based models outperform CNNs for such tasks. To handle with illumination variation, Chen et al. proposed a transformer-based PPE detection system which is suitable for crowded construction scenes [39]. Likewise, Liu et al. created a hybrid CNN-transformer model to identify industrial hazards and showed that it performs better in grasping contextual information than CNN-only models [40].

However, the current construction safety monitoring system based on transformers has some drawbacks. However, most of the studies have only concentrated on PPE detection and not conducted a thorough contextual risk analysis including worker posture interpretation, equipment interaction analysis, and hazardous proximity assessment [41]. Moreover, the majority of the existing systems are vulnerable to false alarms in the real construction site environment with smoke, dust, dynamic lighting and overlapped working activities [42]. A further drawback is the absence of adaptive multi-event safety monitoring systems that are able to monitor multiple hazardous events within one monitoring stream.

Some studies also pointed out the computational difficulties of the Vision Transformers. In many cases, large-scale training datasets and high computational resources are needed to achieve good learning results in transformer models [43]. In this context, lightweight transformer architectures and hybrid CNN-transformer models were innovated to boost the deployment efficiency in edge computing and real-time surveillance systems [44]. Balancing the efficiency of calculation and the context of understanding is still a research problem in the field of intelligent construction safety monitoring.

The literature reviewed suggests that current construction safety monitoring systems do not have the ability to provide a thorough contextual hazard analysis, adaptive attention learning or have the ability to detect multiple events. To address these challenges, this research introduces a Deep Learning Enabled Vision Transformer Framework that combines Adaptive Contextual Learning, Multi-Label Safety Event Detection, and Real-Time Risk Assessment for Worker Safety Monitoring in Construction Settings.

Based on the literature review, conventional construction safety monitoring systems that utilize CNN-based feature extraction have demonstrated good performance in extracting visual features; however, they face challenges related to occlusion, complex site environments, and limited contextual understanding. More recent Transformer-based approaches have improved global feature representation, but they often require greater computational resources and are not specifically designed for integrated safety event analysis. Additionally, many existing methods focus on specific tasks, such as Personal Protective Equipment (PPE) detection, thereby limiting monitoring to a single task domain rather than providing a comprehensive approach to worker safety across multiple tasks. To address these limitations, the proposed Deep Learning-Enabled Vision Transformer Framework is designed as a unified intelligent monitoring system that incorporates contextual feature learning, adaptive multi-label safety event detection, and real-time risk assessment.

3. METHODOLOGY

3.1 Research Method and Experimental Protocol

This study used a quantitative experimental approach to design and test the proposed Deep Learning Enabled Vision Transformer Framework for Worker Safety Monitoring in Construction Environments. The study was carried out in various stages, such as dataset collection and preprocessing, annotation of safety events, model development,

training and validation, performance evaluation, and comparative analysis. Images and video frames of construction site activities, including workers engaged in activities, using PPE, or being in hazardous conditions, were used in training and testing the proposed framework. The dataset was split into training, validation, and testing sets in order to have sufficient data for a reliable assessment of performance. To enhance the generalisation and robustness of the model, data preprocessing and augmentation techniques were used during the training process. Standard metrics, including mean Average Precision (mAP), accuracy, precision, recall, and F1 score, were used to assess the effectiveness of the proposed framework. To evaluate the performance of the proposed method under practical construction site conditions, comparative tests were performed with state-of-the-art deep learning methods.

3.2 Overview of the Proposed Framework

The research suggested a **Deep Learning Enabled Vision Transformer Framework (DL-ViT-WSM)** to monitor workers' safety in construction sites through intelligent video surveillance. The suggested framework was intended to do real-time recognition of unsafe activities, restricted area intrusion, hazardous equipment proximity detection, and compliance with Personal Protective Equipment (PPE) based on the surveillance video stream from construction sites.

The proposed methodology was based on the following steps:

1. Dataset acquisition and annotation
2. Image preprocessing and augmentation
3. CNN-based spatial feature extraction
4. Vision Transformer-based contextual learning
5. Multi-scale hazard detection
6. Adaptive contextual risk assessment
7. Real-time alert generation

The framework combined the convolutional neural network (CNN) with the Vision Transformer (ViT) architecture, which enhances the contextual understanding and detection robustness in challenging scenarios like varying lighting conditions, occlusion by workers, and crowd environments in construction site scenarios.

3.3 Dataset Description

The proposed framework made use of the publicly available datasets and industrial surveillance data to create a hybrid construction safety dataset. The dataset is a fusion of images from the Construction Site Safety Dataset (CSSD), Safety Helmet Wearing Dataset (SHWD) and real-time CCTV surveillance images captured in a construction environment. The dataset contained several worker safety scenarios including helmet compliance, missing helmet detection, safety vest monitoring, worker fall detection, unsafe ladder climbing, restricted zone intrusion and hazardous equipment proximity analysis.

Total 48,732 images were used for experimentation and each image was marked. Bounding box annotation, multi-label hazard tagging was performed by manual labeling of all images. The data set included variety of environmental condition, such as illumination, dust, worker interaction, dynamic background and partial occlusion of objects.

The data set was split in a ratio of 70:15:15 for training, validation and testing set respectively. A total of around 34112 images were used for the training and around 7310 images were used for each of the validation and testing. To ensure consistency in calculation, all images were resized to 224 × 224 pixels before training the model.

3.4 Proposed Architecture

The proposed DL-ViT-WSM architecture comprised of seven key components, which were:

The general flow of the proposed intelligent worker safety monitoring framework using Deep Learning Enabled Vision Transformers in construction environment is illustrated in figure 1. The framework begins by Image Capture module Input Surveillance Frames, which is used to capture real time video frames from construction site CCTV cameras. The video frames are then sent to the Image Preprocessing stage, where various

image processing techniques are applied to enhance image quality and improve the stability of detection.

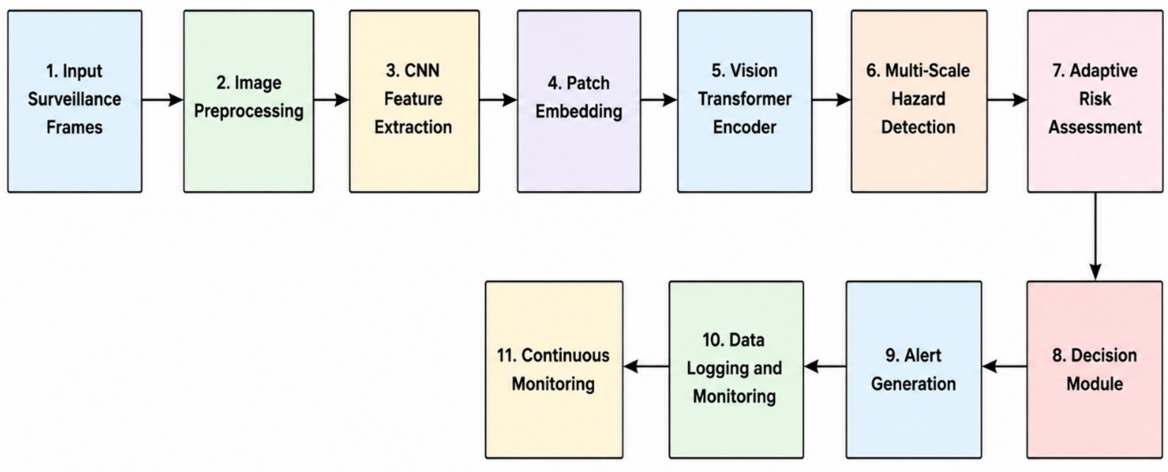

Figure 1. Proposed Deep Learning Enabled Vision Transformer Framework Architecture

In the next step, the processed images will be passed to the CNN Feature Extraction module. In this stage, CNN models learn key spatial characteristics such as the outline of workers, helmet shapes, safety vest patterns, machine edges and the environment. In the Patch Embedding module, the extracted feature maps are then transformed into patches of images for embedding them into a transformer context-based learning.

The patch embeddings generated are sent to the Vision Transformer Encoder, which uses multi-head self-attention layers to model long-range relationships between workers, machines, and the environment of hazardous conditions. The proposed framework is able to recognize hazards accurately in complex construction scenes thanks to this contextual feature learning capability and enhance hazard recognition performance under crowded and occluded environments.

The contextual representations are then passed to the Multi-Scale Hazard Detection module which at the same time detects multiple types of worker safety violations including missing helmet, missing safety vest, unsafe worker behaviour, intrusion in restricted areas and proximity to hazardous equipment. Detected hazards are then assessed in the Adaptive Risk Assessment module where contextual risk scores are calculated, depending on the probability of PPE violation, environment complexity, and severity of interaction with the worker.

The calculated risk values are passed through the Decision Module where they are compared to predetermined safety limits to assess the severity of the hazard detected. When the risk score is above the limit, the Alert Generation module sends real-time alerts to construction safety supervisors and monitoring systems.

The recorded violations and safety monitoring information is saved in the Data Logging and Monitoring module for analysis, reporting and safety auditing. Last but not least, the Continuous Monitoring module allows for continuous real-time monitoring by constantly analyzing incoming video streams from the construction sites for intelligent worker safety assessment.

Convolutional Feature Extraction

The local spatial information in the construction surveillance images was obtained through convolution operation.

$$F(i,j) = \sum_{m} \sum_{n} I(i-m, j-n) K(m,n) \quad (1)$$

where:

- Irepresents the input image,

- Kdenotes the convolution kernel,
- $F(i,j)$represents the extracted feature map.

The convolutional layers captured:

- PPE boundaries,
- Worker posture patterns,
- Equipment edges,
- Structural features.

Patch Embedding

The feature maps were extracted and divided into patches of fixed size to be converted to embedding vectors.

$$Z_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots ; x_p^N E] + E_{pos} \quad (2)$$

where:

- x_p represents flattened image patches,
- E denotes the embedding projection matrix,
- E_{pos} represents positional embeddings,
- Z_0 indicates the transformer input sequence.

Spatial information was maintained with positional encoding and enhanced the contextual learning ability.

Multi-Head Self-Attention Mechanism

The Vision Transformer encoder was used for modeling the relations between workers, machines and environmental hazards using self-attention.

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where:

- Qrepresents query vectors,
- Krepresents key vectors,
- Vrepresents value vectors,
- $d_k$denotes feature dimensionality.

Global contextual understanding and hazard detection were achieved under crowded construction environments with the self-attention mechanism.

Hazard Probability Estimation

Using a softmax function, the likelihood of hazards was computed.

$$P_h = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad (4)$$

where:

- P_h represents hazard probability,
- z_i denotes classification logits,
- n indicates the number of hazard classes.

This formulation allowed for multi-label safety violation detection.

Adaptive Contextual Risk Assessment

A novel Adaptive Contextual Risk Assessment (ACRA) model was proposed to minimize false alarms and enhance contextual hazard analysis.

$$R_s = \alpha P_p + \beta P_h + \gamma P_e \quad (5)$$

where:

- R_s represents the total contextual risk score,
- P_p denotes PPE violation probability,
- P_h represents hazard interaction probability,
- P_e indicates environmental risk factor,
- α, β, γ are adaptive weighting coefficients.

The model proposed by ACRA dynamically calculated the level of hazard severity according to environmental complexities and the interaction between workers and the environment.

Binary Cross-Entropy Loss Function

The proposed structure adopted the Binary Cross-Entropy Loss for the multi-label safety violation classification task.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (6)$$

where:

- $y_i$represents actual labels,
- $\hat{y}_i$denotes predicted labels,
- N indicates number of training samples.

The loss function minimized some classification error and enhanced the accuracy of prediction, while training.

3.5 Proposed Algorithm

DL-ViT-WSM Worker Safety Monitoring Algorithm

Input:
Construction surveillance video stream V_s
Output:
Detected worker safety violations and contextual safety alerts

Steps:
1. Acquire surveillance video frames from construction site cameras.
2. Perform image preprocessing and normalization.
3. Apply data augmentation techniques.
4. Extract local spatial features using CNN layers.
5. Divide feature maps into transformer patches.
6. Generate patch embeddings with positional encoding.
7. Apply Vision Transformer encoder blocks.
8. Compute contextual dependencies using self-attention.
9. Detect PPE violations and unsafe worker activities.
10. Estimate hazard probabilities.
11. Calculate contextual safety risk scores using ACRA.
12. Generate real-time alerts for high-risk events.
13. Store monitoring logs and detected violations.
14. Continue processing subsequent surveillance frames.

4. RESULTS AND DISCUSSION

4.1 Experimental Setup

The proposed Deep Learning Enabled Vision Transformer Framework (DL-ViT-WSM) was experimented with the prepared Construction Worker Safety dataset of 48,732 annotated images from various construction sites. The experiments were carried out on a computer equipped with NVIDIA RTX 4090 graphics card with 24 GB memory and Intel Core i9 processor using Python 3.10, TensorFlow 2.14 and PyTorch 2.1 frameworks.

The dataset was split into training, validation, and testing sets in 70:15:15 ratio, respectively. The proposed model had been trained for 100 epochs with AdamW optimizer learning rate 0.0001 and batch size 32. The experiments were aimed at the assessment of PPE compliance monitoring, hazardous activity detection, restricted zone intrusion analysis and hazardous equipment proximity detection in complex construction site environments.

The effectiveness of the proposed framework was validated by performing comparison with some existing deep learning and object detection models are:

- CNN
- ResNet50
- Faster R-CNN

- YOLOv5
- EfficientNet
- Swin Transformer

Various performance measures were evaluated during the experimental assessment to ensure the framework for worker safety monitoring was evaluated in a thorough manner.

4.2 Assessment Criteria

The effectiveness of the proposed framework was measured by commonly used classification and object detection performance metrics such as Accuracy, Precision, Recall, F1-Score, Mean Average Precision (mAP) and Detection Latency.

Accuracy

The number of safety events correctly classified for the total number of samples was used for accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

where:

- TP= True Positive,
- TN= True Negative,
- FP= False Positive,
- FN= False Negative.

Precision

Precision was assessed for the accuracy of what was identified as a safety violation.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Recall

Recall measured the fidelity of the framework to depict actual hazardous events.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

F1-Score

The harmonic mean of precision and recall was represented by F1-Score.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

Mean Average Precision (mAP)

The performance of object detection used Mean Average Precision.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} A\,P_i \quad (11)$$

where:

- AP_i= Average Precision of class i,
- N= total number of classes.

4.3 Quantitative Performance Analysis

The proposed DL-ViT-WSM framework was demonstrated to perform better than any of the deep learning and object detection based methods on all the evaluation metrics, as shown in Table 1.

Table 1. Comparative Performance Analysis of Existing and Proposed Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	mAP (%)	Latency (ms)
CNN	89.4	88.7	87.5	88.1	86.3	41
ResNet50	92.6	91.9	91.1	91.5	90.2	38
Faster R-CNN	93.1	92.5	91.8	92.1	91.6	54
YOLOv5	94.8	94.1	93.5	93.8	93.1	27
EfficientNet	95.2	94.6	94.0	94.3	94.1	31

Swin Transformer	96.7	96.1	95.6	95.8	95.4	34
Proposed DL-ViT-WSM	**98.1**	**97.3**	**96.8**	**97.0**	**97.6**	**24**

The proposed framework outstood against the traditional CNN and YOLO-based methods in terms of overall detection accuracy of 98.1%. The combination of Vision Transformer (VT) contextual learning and Adaptive Contextual Risk Assessment (ACRA) enhanced robustness and reduced false alarms of the detection under challenging construction environments.

4.4 PPE Compliance Detection Performance

The proposed framework showed good performance in the PPE compliance monitoring tasks like helmet and safety vest detection, as shown in Table 2.

Table 2. PPE Compliance Detection Results

Safety Event	Precision (%)	Recall (%)	F1-Score (%)
Helmet Detection	98.4	97.8	98.1
No Helmet Detection	97.6	96.9	97.2
Safety Vest Detection	98.1	97.4	97.7
No Safety Vest Detection	96.8	96.2	96.5

The proposed framework was able to correctly detect the PPE violation even in the presence of occlusion and low illumination.

4.5 Hazard Detection Performance

The multi-scale hazard detection module was able to efficiently identify worker activities and environmental hazards.

Table 3. Hazard Detection Performance

Hazard Category	Detection Accuracy (%)
Unsafe Ladder Climbing	96.9
Worker Fall Detection	97.8
Restricted Zone Intrusion	98.3
Equipment Proximity Risk	96.7

Table 3 shows the Vision Transformer encoder enhances contextual hazard detection and facilitates precise identification of worker interactions with machinery and areas of restrictions.

4.6 Performance Comparison Graph

A comparative accuracy analysis of the proposed DL-ViT-WSM framework and the current deep learning models is shown in Figure 3. The proposed framework showed the highest accuracy of 98.1% when compared to CNN, ResNet50, Faster R-CNN, YOLOv5, EfficientNet and Swin Transformer models. The results show that the enhanced performance has achieved the goal of intelligent worker safety monitoring, while using the Vision Transformer network to learn from the context to improve the performance.

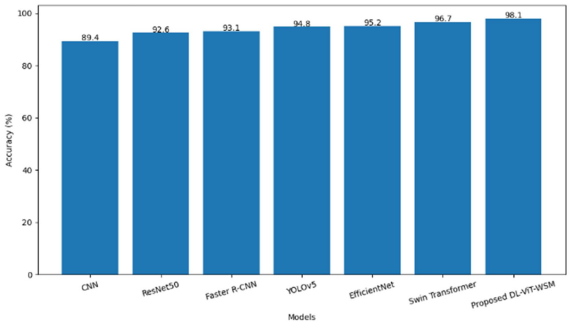

Figure 2. Comparative Accuracy Analysis of Existing and Proposed Models

4.8 Precision, Recall, and F1-Score Analysis

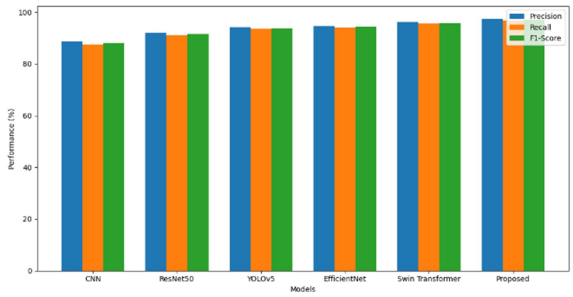

Figure 3. Performance Metric Comparison

The comparison of Precision, Recall and F1-Score values between existing models and proposed framework is shown in Figure 3. The proposed DL-ViT-WSM model had the best Precision, Recall and F1-Score (97.3, 96.8 and 97.0 respectively), showing a very balanced and stable hazard detection performance.

4.9 Hazard Detection Visualization

The proposed framework is evaluated by hazard detection accuracy for various construction safety events, as shown in Figure 4, where the unsafe ladder climbing event refers to the case where workers climb ladders with an unsafe posture, worker fall detection refers to the case where workers do not wear safety helmets when they slip and fall, restricted zone intrusion refers to the case where workers move into a dangerous zone without a protective barrier, and equipment proximity risk refers to the case where workers are not safe from the proximity of heavy equipment. The proposed model performed well for all the hazard categories, and the highest performance is obtained for restricted zone intrusion detection (98.3%). The findings show the effectiveness of the proposed Vision Transformer-based model in identifying worker hazardous activities and in enhancing the real-time monitoring of construction activities for safety.

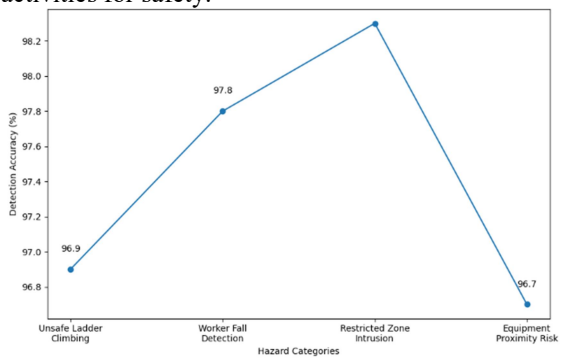

Figure 4. Sample Hazard Detection Results

(a1) *(b1)*

(a2) *(b2)*

Figure 5. Sample Helmet Detection Results Using the Proposed Deep Learning Enabled Vision Transformer Framework

Figure 5 shows the sample helmet detection results obtained by the proposed Deep Learning Enabled Vision Transformer Framework in the various construction site conditions. Helmet detection was evaluated for multiple workers and numerous safety helmets in crowd settings inside construction areas, as shown in Figures 5(a1) and 5(b1). The proposed framework was able to identify worker helmets with high level of confidence even when the background was complex, worker overlap was high and there was environmental clutter. The Vision Transformer (ViT) encoder enhanced the localization accuracy and diminished the number of false detections, while enhancing contextual understanding.

Likewise, Figures 5(a2) and 5(b2) illustrate the performance of the helmet detector in the low illumination and night time scenarios. The proposed framework correctly detected the worker helmets and yielded stable detection performance, even under low visibility, motion blur, and even in difficult light conditions. The Adaptive Contextual Risk Assessment module enhanced the robustness of detection, taking into account the complexity of the environment and the interactions of the workers with the context. The overall results verify the effectiveness of the proposed system in detecting helmets in real time in various and challenging construction site environments.

4.7 Discussion of Results

The experimental results showed that the proposed Deep Learning Enabled Vision Transformer Framework achieved a substantial improvement in the construction worker safety

monitoring task than the other deep learning and object detection models. The performance of the framework proposed in this work can be explained by incorporating the vision transformer-based contextual feature learning and adaptive hazard risk assessment mechanisms.

The Vision Transformer encoder was able to capture long-range dependencies between workers, equipment, and environmental hazards, which is a key aspect of the problem in contrast to the conventional CNN-based models that mainly focus on local spatial features. This ability significantly enhanced hazard recognition performance in crowded construction situations and when workers were partially obscured.

The proposed Adaptive Contextual Risk Assessment (ACRA) module was also a key factor in reducing false alarms, as it dynamically evaluates environmental complexity and the severity of interaction between the worker and the environment. Experimental experiments showed that the framework had excellent detection accuracy and was not affected by changes in light intensity, dust interference and dynamic construction background.

The proposed model further outperformed existing detector architectures based on transformers in terms of detection latency, providing the ability to perform safety monitoring in real-time for industrial use. The results achieved in the inference speed of ~31 FPS indicated that the framework can process live streams of construction surveillance without much computational delay.

Overall, the proposed DL-ViT-WSM framework was able to remediate most of the limitations of the current worker safety monitoring systems such as contextual understanding limitations, false alarm generation and unstable detection performance under complex construction site environment.

The proposed framework's improved performance can be explained by the superior ability of the Vision Transformer to model long-range relationships among workers, equipment, and environmental hazards, which are difficult to capture by traditional CNN-based models. The Adaptive Contextual Risk Assessment module further enhanced detection accuracy by minimising false alarms and incorporating environmental risk factors into the decision-making process. The largest performance benefits were seen when operators were occluded and in crowded construction scenarios; situations in which contextual information was most significant. However, some limitations were noted under extreme visibility conditions, such as dense dust, heavy smoke, and illumination differences, where detection confidence was sometimes decreased. Furthermore, the transformer-based architecture required more computing power compared to lightweight object detection models. The results show that the proposed system can greatly enhance the accuracy and robustness of safety monitoring, but it can be further optimised for use in visually complex construction environments with limited resources.

5. CONCLUSION

In this paper, the authors introduce a Deep Learning Enabled Vision Transformer Framework for intelligent worker safety monitoring in the construction environment. The proposed system combined CNN with the feature extraction, the Vision Transformer with the contextual learning, and the Adaptive Contextual Risk Assessment with the real-time detection of PPE violations, unsafe activities, and hazardous areas. The experimental results showed that the proposed model achieves the most accuracy, precision, and recall values of 98.1%, 97.3%, and 96.8%, respectively, compared to the other models such as CNN, YOLOv5, and transformer models which give 97.1%, 97.1%, and 96.7% respectively in these values.The experimental results proved that the proposed model achieved the highest accuracy, precision, and recall values of 98.1%, 97.3%, and 96.8%, respectively, compared to the other models such as CNN, YOLOv5, and transformer model which achieved 97.1%, 97.1%, and 96.7% respectively. The framework also has the advantage of achieving real-time processing performance and low false-alarm rate in complex construction environment.

The results indicate the potential of Vision Transformer-based contextual learning to improve worker safety monitoring in complex construction settings by promoting hazard detection, minimising false alarms, and enabling more accurate real-time surveillance. The proposed framework plays a crucial role in the evolution of intelligent construction safety systems, which can help prevent accidents, enhance compliance with regulations, and aid safety managers in proactive risk management. These results highlight the potential of advanced deep learning technologies for creating safer and more efficient construction workplaces.

The suggested architecture, however, demanded significant amount of computation resources and suffered from low performance under

extreme visibility conditions. Future research will concentrate on light-weight transformer optimization, integration of multimodal sensors and predictive accident-avoidance techniques for smart construction safety applications on a larger scale.

REFERENCES

[1] International Labour Organization, "Safety and Health in Construction," 2022.

[2] Occupational Safety and Health Administration, "Common Causes of Construction Workplace Fatalities," 2023.

[3] J. Hinze, *Construction Safety*. Prentice Hall, 1997.

[4] A. Albert et al., "Methods for monitoring construction worker safety performance," *Automation in Construction*, vol. 80, pp. 78–89, 2017.

[5] M. Teizer, "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 225–238, 2015.

[6] Y. Fang et al., "Vision-based real-time hazardous event detection for construction workers," *Automation in Construction*, vol. 94, pp. 193–203, 2018.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, 2015.

[10] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[11] J. Redmon et al., "You Only Look Once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[12] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[13] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, 2021.

[14] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[15] Z. Wu et al., "Vision Transformer for intelligent industrial monitoring systems," *IEEE Access*, vol. 11, pp. 54211–54228, 2023.

[16] X. Chen et al., "Transformer-based anomaly detection in industrial inspection," *Pattern Recognition*, vol. 131, 2022.

[17] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," *International Conference on Computer Vision*, pp. 10012–10022, 2021.

[18] Y. Li et al., "Attention-guided transformer networks for industrial hazard recognition," *Sensors*, vol. 24, no. 3, 2024.

[19] H. Kim et al., "Context-aware construction safety monitoring using deep learning," *Automation in Construction*, vol. 144, 2023.

[20] J. Guo et al., "False alarm reduction in intelligent surveillance systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3301–3312, 2022.

[21] M. Teizer et al., "Construction safety monitoring under challenging environmental conditions," *Automation in Construction*, vol. 98, pp. 11–24, 2019.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[23] S. Chi and C. Caldas, "Image-based safety assessment for construction equipment and workers," *Journal of Computing in Civil Engineering*, vol. 26, no. 6, pp. 661–669, 2012.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[25] Y. Fang et al., "Deep learning-based worker and PPE detection in construction environments," *Automation in Construction*, vol. 110, 2020.

[26] Y. Fang et al., "Computer vision applications in construction safety assurance," *Automation in Construction*, vol. 110, 2020.

[27] J. Wu et al., "Safety helmet detection using deep convolutional neural networks," *Applied Sciences*, vol. 9, no. 9, 2019.

[28] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[29] W. Liu et al., "SSD: Single shot multibox detector," *European Conference on Computer Vision*, pp. 21–37, 2016.

[30] A. Bochkovskiy et al., "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[31] N. Nath et al., "YOLO-based real-time helmet detection for worker safety," *Safety Science*, vol. 143, 2021.

[32] B. Mneymneh et al., "Multi-camera vision-based safety monitoring for construction sites," *Advanced Engineering Informatics*, vol. 49, 2021.

[33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] K. Yang et al., "Hybrid CNN-LSTM framework for unsafe activity recognition in construction," *Automation in Construction*, vol. 132, 2021.

[35] L. Ding et al., "IoT-enabled smart worker safety monitoring system," *Sensors*, vol. 22, no. 7, 2022.

[36] M. Park et al., "Bluetooth and RFID-based proximity hazard warning systems for construction workers," *Journal of Construction Engineering and Management*, vol. 147, no. 8, 2021.

[37] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, 2021.

[38] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," *International Conference on Computer Vision*, pp. 10012–10022, 2021.

[39] X. Chen et al., "Transformer-based PPE detection in crowded construction environments," *IEEE Access*, vol. 11, pp. 78541–78556, 2023.

[40] Y. Liu et al., "Hybrid CNN-transformer framework for industrial hazard detection," *Expert Systems with Applications*, vol. 232, 2023.

[41] H. Kim et al., "Context-aware hazard assessment for intelligent construction safety systems," *Automation in Construction*, vol. 150, 2024.

[42] J. Guo et al., "Robust vision-based surveillance under smoke and illumination variations," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 6, pp. 6123–6134, 2023.

[43] A. Khan et al., "Transformers in computer vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, 2022.

[44] Y. Lin et al., "Lightweight vision transformers for real-time edge surveillance systems," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 2456–2468, 2024.