International Journal of Safety and Security Engineering
Vol. 15, No. 5, May, 2025, pp. 859-871
Journal homepage: http://iieta.org/journals/ijsse

Automated Detection of Construction Opening Safety Risks Using Computer Vision

Yih-Tzoo Chen, Ting-Chuan Hu*

Department of Construction Engineering, National Kaohsiung University of Science and Technology, Kaohsiung 82445, Taiwan

Corresponding Author Email: F112106118@nkust.edu.tw

https://doi.org/10.18280/ijsse.150501

Received: 14 April 2025
Revised: 16 May 2025
Accepted: 25 May 2025
Available online: 31 May 2025

Keywords:
computer vision, construction safety, YOLO model, image-text semantic matching, fall prevention

ABSTRACT

High fall accidents are frequently reported on construction sites, with inadequate opening protection being one of the major risk sources. Traditional manual inspection methods are inefficient and susceptible to subjective judgment. To enhance the efficiency and accuracy of Construction Safety Management (CSM), this study aims to develop an AI-based automated detection and report generation system. The system integrates YOLOv8 for object detection and the CLIP model for image-text matching, enabling real-time analysis of construction site images to automatically identify the absence of fall protection or non-compliant installations and generate structured safety reports. The YOLOv8l-based object detection model yielded a precision of 0.943, recall of 0.952, and mAP@50 of 0.923. The CLIP-based semantic matching module achieved an accuracy of 85% for identifying non-compliant fall protection conditions at an optimized similarity threshold of 0.66. These outcomes support the system's risk detection and regulatory compliance verification effectiveness. This system not only improves inspection efficiency and standardization of reports but also reduces human error, strengthens real-time hazard identification and response capabilities, and showcases its potential in driving the digital transformation of construction safety management.

1. INTRODUCTION

The construction industry is one of the least digitized sectors globally and, due to its highly labor-intensive nature, is also a high-risk industry where worker injuries frequently occur. Common causes of injuries include improper working postures, manual handling of heavy objects, and high-intensity operations [1]. Among these, Falls from Heights (FFH) are considered one of the primary causes of severe injuries and fatal accidents on construction sites. Studies showed that falls from heights account for 48% of all construction-related accidents, with scaffolding and roofing operations posing particularly high risks [2].

A study focusing on construction sites in Hong Kong analyzed the root causes of fall-from-height incidents and found that a lack of adequate protective equipment at open edges often led to accidents during high-elevation work. The study suggested that improving site design and providing better safety measures could effectively reduce the risk of falls [3]. Additionally, research highlighted that falls from heights represented a significant portion of construction accidents, underscoring the importance of standardized fall protection systems, especially in edge work operations [4].

The impact of fall-from-height accidents not only results in immediate injuries but also brings about substantial social and economic losses, particularly in cases where workers fall from unprotected edges or heights. In response, some studies proposed enhancing workers' safety awareness and strictly enforcing the correct use of protective equipment. These measures were considered crucial for effectively reducing the occurrence of such accidents [5].

With the rapid development of Artificial Intelligence (AI), advanced technologies such as Computer Vision (CV) and Natural Language Processing (NLP) were progressively applied to the field of construction management, aiming to assist safety personnel in monitoring construction site safety [6-9]. The high accident rate in the construction industry caused significant social harm, making the effective implementation of safety management a key focus. For instance, Fang et al. [10] developed an Image-Text Semantic Matching (ITSM) model to evaluate the semantic similarity between visual and textual features, determining whether workers' behaviors in images violated the safety regulations recorded in text. However, the model's binary output format limited scalability. Ding et al. [11] proposed a Visual Question Answering (VQA)-based inference framework for training and application in safety compliance checks. Emerging information technologies such as the Internet of Things (IoT) and computer vision were also applied to on-site safety monitoring [12]. These technologies require recording and storing large amounts of data, such as digital images and incident reports, as a basis for preventing future accidents [13]. However, extracting insights from massive datasets for safety management and decision-making remained a challenge [14].

Research showed that enhancing the standardization and transparency of reports helped strengthen communication

effectiveness and consistency within and outside organizations. Standardized reports provided a reference framework, simplified reporting processes, and ensured the accuracy and completeness of information transmission [15]. Furthermore, applying digital and intelligent technologies to optimize reporting processes through data management further enhances decision-making efficiency and information transparency. This was particularly important in construction management, where digital reporting improved information consistency and simplified the overall process [3]. Specifically, customized reporting processes allowed digital architectures to adjust according to the needs of multiple stakeholders, ensuring that reports accurately met various demands, especially in engineering projects involving multi-party collaboration [16]. Overall, enhancing standardization and transparency made reporting processes more consistent across different business scenarios, enabling more efficient and consistent satisfaction of diverse management requirements in construction project management [15].

Although recent studies have increasingly applied artificial intelligence to construction safety management and have developed image recognition–based risk detection systems, two major limitations remain. First, most existing image-text semantic matching (ITSM) approaches focus only on single-level semantic alignment, often overlooking subtle distinctions related to the correctness of installation and regulatory compliance of fall protection facilities around openings. This results in limited precision and insufficient semantic depth in the detection outcomes. Second, current systems emphasize violation detection but lack mechanisms for tracking subsequent corrective actions, making it challenging to implement closed-loop risk management and verification.

To address these issues, this study proposes an automated detection system that integrates YOLO for object detection and CLIP (Contrastive Language Image Pretraining) for image-text semantic matching. The system improves semantic resolution and detection accuracy while incorporating a report generation and feedback module to support real-time monitoring and continuous risk tracking. This integrated approach addresses the shortcomings of existing AI-based safety systems by enabling both precise risk identification and closed-loop safety management.

2. RELATED WORK

2.1 Construction safety management (CSM)

Construction safety management (CSM) comprises a set of strategies, practices, and methodologies specifically designed to ensure the safety of construction labor forces. Its primary objective is to systematically predict, identify, and prevent risks that might endanger construction workers, public safety, or infrastructure integrity [17]. The construction site's temporary and dynamic nature and frequent interactions among personnel, machinery, and materials created an environment prone to unforeseen hazards. Common incidents, such as "falls from height" and "being struck by objects," not only highlighted the frequency of accidents but also revealed systemic challenges within existing safety regulations [18]. Many accidents occurred at scaffolding, rooftops, and open edges, particularly when adequate protective equipment was lacking or structural supports were unstable. Preventing such incidents required rigorous equipment inspections and a stable construction environment [2, 19]. Additionally, failures in management systems, insufficient supervision, and inadequate safety management plans further exacerbated the risks of falls. The absence of standardized management processes and effective safety education was also identified as a root cause of accidents [20].

Artificial Intelligence (AI) technologies made significant contributions to construction safety, encompassing applications in machine learning (ML), computer vision (CV), natural language processing (NLP), knowledge-based systems, and robotics [21]. Computer vision, deep learning, and action recognition were among the most commonly utilized technologies in construction safety, with primary applications in risk assessment and real-time monitoring. Research highlighted that AI, particularly computer vision, demonstrated notable success in construction safety. For instance, computer vision detected whether workers were wearing personal protective equipment (PPE) such as helmets or harnesses, thereby reducing the likelihood of falls from height and other accidents. Moreover, computer vision technologies identified potential hazards on construction sites, such as unprotected edges and unstable scaffolding, and enhanced safety through real-time monitoring [22]. Although the research topics in construction safety were diverse, they could be divided into two perspectives: management-driven and technology-driven. The first perspective suggested that improving management performance ensured construction safety and prevented site accidents. This research direction typically included safety climate, safety culture, worker skills and behaviors, and risk management. The second perspective focused on using various technologies to ensure construction site safety, including developing new technological tools to promote real-time monitoring and risk prevention [17].

Table 1. Various computer vision tasks are used in safety management

Task	Primary Applications	Major Algorithms/Models
Object recognition	Detect construction site openings, guardrails, safety nets, and covers	YOLOv8, YOLOv10, Faster R-CNN
Object tracking	Tracking worker movements, machinery operations	DeepSORT, ByteTrack, SORT
Action recognition	Detecting unsafe behaviors (e.g., not wearing safety equipment)	TimeSformer, PoseC3D, RNN
Image-text semantic matching	Automatic comparison with construction regulations, safety violation analysis	CLIP, ALIGN, BLIP

Construction safety management aims to protect workers from various potential risks through systematic strategies and practices, and the rise of AI technologies has further reinforced this goal. Technologies such as computer vision and deep learning enable real-time monitoring of site safety conditions,

helping to identify and reduce the occurrence of hazardous events, such as falls from height. Meanwhile, integrating management-driven and technology-driven perspectives contributes to more efficient safety management, enhancing overall construction site safety and management effectiveness.

In the field of fall risk prevention in construction, researchers proposed various methods to mitigate the risks associated with working at heights while also addressing the limitations of these approaches. First, risk identification and assessment based on risk matrices was a standard method that categorized hazards during construction processes to develop targeted preventive measures. However, this method relied heavily on managers' understanding and subjective judgment of the construction site, which might have overlooked latent risks. Furthermore, the effectiveness of its preventive measures was limited by the decision-makers' experience [23].

2.2 Computer vision

2.2.1 Object Recognition models

The core tasks of computer vision encompass various functionalities, including image classification, object detection, image segmentation, and visual question answering. These tasks aimed to enable machines to understand and analyze visual data. Recent advancements in deep learning technologies have significantly improved the performance of these tasks [24, 25]. Computer vision provides machines with human-like visual capabilities. Through specialized algorithms, it processes visual information, identifies objects, understands contexts, and detects anomalies. Table 1 summarizes common computer vision tasks and some widely adopted algorithms used in safety applications.

The main objective of this study is to ensure that fall protection facilities around construction site openings can be effectively detected and further verify whether their installation complies with safety regulations. Therefore, this study will delve into the technical applications of Object Recognition and image-text semantic matching models.

In the Object Recognition section, this study aims to automate the identification of protective facilities around construction site openings, including guardrails, fences, and temporary protective structures, through deep learning techniques. Traditional manual inspection methods are time-consuming and prone to subjective judgment, while the introduction of Object Recognition technology significantly improves detection accuracy and efficiency, reducing human error.

Common Object Recognition models are listed in Table 2. It can be observed that the YOLO series models show significant advantages in processing speed compared to other models. Due to its high speed and single-stage architecture, the YOLO (You Only Look Once) model is widely adopted in various object detection applications. With continuous optimization over versions, YOLO models effectively addressed challenges in diverse application scenarios [26].

Table 2. Object Recognition model performance comparison

Architecture	mAP@50	GPU Latency
YOLOv8	0.62	1.3ms
Faster R-CNN	0.41	54ms
EfficientDet	0.47	N/A

Some studies proposed a globally optimized YOLO-based object detection technique for construction site applications that emphasized improving detection performance in dynamic environments. By integrating spatial interaction information, the model's overall detection capability in construction environments was significantly enhanced [27]. Another study compared the performance of YOLOv5 and YOLOv8 in detecting construction risk factors. The results showed that both versions achieved excellent accuracy in recognizing workers and personal protective equipment (PPE) [28].

Regarding high-risk areas, one study applied the YOLO model to detect open edges. It utilized a specially annotated dataset for training, which enabled the model to identify hazardous boundaries at construction sites accurately and instantly alert workers to enhance on-site safety [29]. Additionally, in construction safety management, the YOLO model was employed to monitor heavy equipment and worker behavior, such as detecting whether workers had entered hazardous zones or failed to comply with safety regulations, thereby improving overall construction management efficiency [30].

Therefore, this study selects the YOLO series model to detect the presence of fall protection facilities around site openings. The system can learn and recognize various types of fall protection facilities by collecting many real-world construction site images for training. The system automatically generates alerts when it detects the absence of protection or improper installation around openings.

2.2.2 ITSM models and methods

In the image-text semantic matching section, this study employs Cross-modal Technology, utilizing models like CLIP (Contrastive Language-Image Pretraining) to semantically compare the protective facilities in images with construction safety standard texts to verify compliance.

Contrastive Learning-Based Models: These models employed contrastive learning to embed images and texts into a shared semantic space. For example, the CLIP model demonstrated outstanding performance in open-domain image retrieval and cross-modal search. The training objective of CLIP was to bring semantically related image-text pairs closer and push unrelated pairs further apart, showing exceptional performance in zero-shot learning where limited annotations were available [31, 32].

Local and Global Matching Models: Models like VisualBERT and UNITER segmented images into multiple regions and aligned them with corresponding text tokens to achieve more refined semantic matching. This method was particularly suitable for application scenarios that required high semantic alignment, such as scene understanding or medical image analysis, where precision was crucial [33].

Multi-View Attention Models: Some models adopted multi-view attention mechanisms to enhance the accuracy of ITSM. For example, the Multi-View Attention Model (MVAM) represented images and texts from different perspectives to capture richer semantic information and improve matching performance. This approach was suitable for semantically complex scenarios, enabling the model to capture subtle differences between images and texts [34].

Among these, CLIP stood out for its diversity and flexibility, making it a foundational model in visual-language processing. It was widely applied in cross-domain ITSM systems, such as cross-modal retrieval and visual question answering tasks. Compared to specialized task-oriented models like multi-view attention and local-global matching models, CLIP’s design was more adaptable to open-domain and unstructured tasks

due to its large-scale image-text paired training, which enhanced its semantic alignment capabilities and cross-domain generalization [35, 36]. Furthermore, CLIP demonstrated exceptional performance in zero-shot learning, cross-modal retrieval, and open-domain understanding tasks [31, 32].

In recent years, YOLO and CLIP models have been widely applied in construction safety, demonstrating significant technological potential. YOLO models, with their real-time inference speed and high detection accuracy, were particularly well-suited for rapid-response risk monitoring on construction sites. By incorporating attention mechanisms and feature fusion strategies, YOLO models effectively detected objects such as safety helmets and guardrails even in complex site environments [37, 38]. However, YOLO still had limitations in detecting small or partially occluded objects and performed inconsistently under extreme lighting or visual interference conditions [39]. The CLIP model, on the other hand, possessed strong cross-modal semantic alignment capabilities, enabling connections between images and text and supporting semantic image-text queries, which showed promising applications in intelligent construction management [40]. Nevertheless, CLIP's limited sensitivity to fine-grained local features restricted its ability to identify minor violations in real construction site scenarios [41].

Based on the literature review in this section, detecting fall protection measures around construction site openings is critical for ensuring worker safety. However, traditional inspection methods rely heavily on manual inspections, which are inefficient and prone to subjective errors. In previous studies, deep learning technologies, particularly Object Detection and image-text semantic matching (ITSM), have demonstrated significant effectiveness and are considered capable of substantially enhancing safety monitoring efficiency on construction sites.

Building upon the validation of these research findings and the successful application of these technologies, this study aims to integrate YOLO and CLIP techniques further. By leveraging YOLO's high-speed object detection capabilities and the cross-modal semantic matching of CLIP, the objective is to develop an automated detection system for construction site opening protection. This system is expected to automatically detect fall protection facilities around construction site openings, perform real-time assessments of their compliance with safety regulations, and ultimately generate standardized inspection reports to enhance the efficiency and accuracy of construction safety management.

3. METHODOLOGY

3.1 Object Recognition model

This section introduces an Object Recognition model for image recognition of construction site openings and fall protection facilities. The objective is to enhance the model's generalization ability through a diverse training dataset. This dataset includes images of construction site openings and fall protection facilities collected from open computer vision platforms. Images are annotated using a randomized method (detailed in Section 3.1.1) and are divided into training, testing, and validation sets. This partition helps fine-tune parameters during training to prevent overfitting while ensuring robust performance on unseen data.

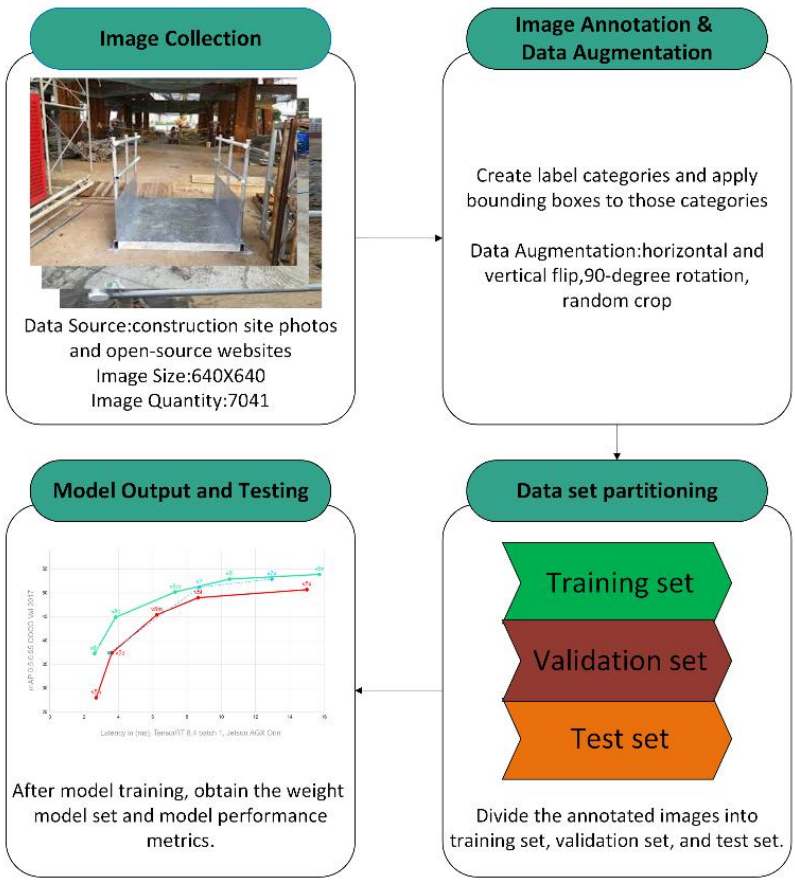

Figure 1. Model training process

To improve recognition performance, the training data undergoes preprocessing and data augmentation. The training process is conducted using the PyTorch framework. Various evaluation metrics are introduced to evaluate the model's performance comprehensively. The framework is illustrated in Figure 1.

3.1.1 Hazard image dataset

The dataset employed in this study primarily consists of images of construction site openings captured on-site, along with images of fall protection facilities—including guardrails, opening covers, and safety nets—collected from open-source computer vision platforms. All images were annotated using labelImg, and the dataset was randomly divided into training, validation, and testing subsets in a ratio of 8:1:1. The training set was used to facilitate model learning, the validation set was utilized during the training phase to monitor performance, adjust hyperparameters, and prevent overfitting, while the testing set was reserved for final evaluation to assess the model's generalization on unseen data. The original dataset included 184 on-site images of construction openings and 2,163 of fall protection facilities. After applying data augmentation techniques, the total number of images in the dataset increased to 7,041.

To enhance the model's generalization capability, a series of preprocessing and data augmentation operations were applied to the dataset. First, all training images were subjected to auto-orientation to ensure consistency in image direction. Second, to meet the input requirements of the neural network, the dimensions of each image were uniformly resized to 640x640 pixels. In terms of data augmentation, to simulate variations in perspectives and scales, the following methods were applied to each training sample, generating three transformed images:

Horizontal and Vertical Flipping: Simulates changes in object orientation within the scene.

90-Degree Rotation: Includes clockwise, counterclockwise rotations, and vertical flipping to enhance the model's robustness to directional changes.

Random Cropping: Randomly crops images within a scaling range of 0% to 20%, helping the model learn to recognize objects from partial perspectives.

3.1.2 Model development and training

This study constructed an object detection model based on YOLOv8 using the PyTorch framework in a Python 3.11 environment within PyCharm. All experiments were conducted on a PC equipped with the following specifications: **CPU**: 12th Gen Intel(R) Core(TM) i7-12700KF 3.60 GHz, **GPU**: NVIDIA GeForce RTX 3070-12G, and **RAM**: 32 GB.

Table 3. Key hyperparameters

Parameters	Value
Model	YOLOv8, YOLOv10
Batch size	8
Epochs	100
Optimizer	AdamW
Loss Function	Cross-Entropy Loss
IoU	0.7
Learning Rate	0.00125

Before training the YOLO model, a pre-trained base model was selected to enhance efficiency and performance. The network architecture was adjusted as needed, and key hyperparameters like learning rate and batch size were optimized (Table 3). During training, the model processed image data to generate predictions, calculated loss values against ground truth, and updated parameters via backpropagation. This iterative process continued until the performance was satisfactory or the set number of epochs was reached.

The table presents a series of key parameters configured during the training of the YOLOv8 and YOLOv10 models. These parameters are carefully selected to optimize the training process and enhance model performance, as explained below:

Model:

This study will test all models from YOLOv8 and YOLOv10.

Batch Size:

Set to 8, representing the number of images processed in each iteration. A large batch size may result in more stable gradient estimation, but can cause memory overflow on limited hardware. Conversely, a small batch size may lead to high gradient variance, affecting the stability of model convergence. Empirical testing revealed that a medium-sized batch provides a good trade-off between training efficiency and resource constraints, ensuring consistent training progress.

Epochs:

Set to 100, indicating that the training data is used in full 100 times. Insufficient epochs may result in underfitting, where the model fails to learn adequate features. However, excessively high epochs may lead to overfitting, particularly without regularization or early stopping mechanisms. Appropriate epoch selection should be guided by the model's validation performance trend and dynamically adjusted based on convergence behavior.

Optimizer:

The AdamW optimizer was employed, which uses adaptive learning rates and is designed to improve model performance by adjusting parameters dynamically during training. Compared to traditional optimizers such as SGD or standard Adam, AdamW improves convergence efficiency in high-dimensional parameter spaces and helps reduce overfitting, which is especially important when training deep neural networks on safety-critical datasets.

Loss Function:

Cross-entropy loss was used, a function commonly applied to classification tasks to measure the difference between the predicted probability distribution and the actual labels.

Intersection over Union (IoU):

Set to 0.7, this metric evaluates the overlap between predicted and ground truth bounding boxes. If set too low, the model may overestimate accuracy, resulting in false positives. If set too high, it may miss partially correct predictions, increasing false negatives.

Learning Rate:

Set to 0.00125, a key factor determining the magnitude of weight updates. A well-chosen learning rate helps the model learn effectively, avoiding oscillations or excessively slow convergence during training. A high learning rate may cause unstable gradients or oscillation, especially in the early training stages. On the other hand, a very low learning rate slows down convergence, preventing the model from reaching optimal performance within a reasonable timeframe. Selecting a moderate and dynamically adjustable learning rate contributed to training stability and convergence efficiency in this study.

3.1.3 Evaluation metrics

In object detection models, commonly used evaluation metrics include:

Intersection-over-Union (IoU):

To evaluate the model's performance, ground truth bounding boxes (manually annotated) are compared with the predicted bounding boxes. The Intersection-over-Union (IoU) metric is calculated as the ratio of the overlapping area between the predicted bounding box and the ground truth bounding box (intersection, ∩) to the combined area of both bounding boxes (union, ∪), as shown in Eq. (1):

$$\boldsymbol{IoU} = \frac{\boldsymbol{Area\ of\ Overlap}}{\boldsymbol{Area\ of\ Union}} = \frac{|\boldsymbol{B_{pred} \cap B_{gt}}|}{|\boldsymbol{B_{pred} \cup B_{gt}}|} \quad (1)$$

where,

B_{pred} denotes the predicted bounding box generated by the object detection model.

B_{gt} denotes the ground truth bounding box annotated in the dataset.

Average precision (AP):

Precision and recall are calculated for various confidence score thresholds, allowing the precision-recall curve to be plotted. AP is computed as the area under the precision-recall curve, as defined in Eq. (2):

$$\boldsymbol{AP} = \int_0^1 \boldsymbol{P(r)dr} \quad (2)$$

where,

P(r) denotes the precision as a function of recall r, describing the model's precision at a given level of recall.

r represents the recall, defined as the ratio of correctly predicted positive instances to all actual positive instances.

Mean average precision(mAP):

To evaluate the model's overall performance, mAP is calculated as the mean of the AP values across all categories. In object detection, mAP is traditionally computed for a single IoU threshold (denoted as mAP_{IoU}). In YOLO series models mAP_{50} *and* mAP_{50-95}, are commonly used to represent precision metrics, as shown in Eq. (3):

$$\boldsymbol{mAP} = \frac{1}{C}\sum_{C=1}^{C} \boldsymbol{AP_C} \quad (3)$$

where,

C is the total number of categories.

$\boldsymbol{AP_c}$ is the average precision for category c.

Accuracy:

Accuracy is the ratio of correct predictions to the total number of predictions, as shown in Eq. (4):

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (4)$$

Precision:

Precision is the ratio of true positive predictions to the total number of positive predictions, as shown in Eq. (5):

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (5)$$

Recall:

Recall is the ratio of true positive predictions to the total number of actual positive cases, as shown in Eq. (6):

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (6)$$

F1-score:

The F1-score combines precision and recall into a single metric, calculated as their harmonic mean, as shown in Eq. (7).

$$\boldsymbol{F1-Score} = \boldsymbol{2} * \frac{\boldsymbol{Precision * Re\ call}}{\boldsymbol{Precision + Re\ call}} \quad (7)$$

Notes:

True Positive (TP): The number of positive samples correctly predicted as positive.

False Positive (FP): The number of negative samples incorrectly predicted as positive.

True Negative (TN): The number of negative samples correctly predicted as negative.

False Negative (FN): The number of positive samples incorrectly predicted as negative.

3.2 Image-text semantic matching

This section explains the prepared textual descriptions, including installation requirements and safety regulations for fall protection facilities, which serve as the basis for image comparison. It also details the threshold settings to ensure the model accurately filters out non-compliant images. Additionally, the image-to-semantic matching process is described as enhancing detection accuracy.

3.2.1 Text preparation

This study aims to address safety management issues on construction sites, particularly focusing on identifying openings and fall protection facilities. To achieve this, content related to fall hazard prevention facilities was extracted from the *Standards for Construction Safety and Health Facilities*. Since the original legal texts are often lengthy, they were transformed into a format more suitable for CLIP model analysis. This process involved simplifying the detailed descriptions in the regulations into shorter sentences, enabling the model to perform ITSM more effectively.

Research indicated that the CLIP model performed more efficiently in learning and predicting image content when processing concise and precise textual descriptions [42].

Data collected from social media platforms further demonstrated how short texts improved image-text alignment efficiency in various contexts, especially descriptive text and image pairing [43]. A novel contrastive training method introduced in recent studies optimized representation alignment through large-scale image-caption and text-text pairs, allowing the model to excel in text-image and text-text matching tasks. This highlighted the importance of concise sentences in enhancing the performance of the CLIP model [44]. The training and application of the CLIP model consistently demonstrated its superior ITSM capabilities, particularly in zero-shot learning scenarios when dealing with new images similar to the training data distribution [42]. Noted that the CLIP model, through its contrastive learning framework, effectively aligns image and text representations, enabling accurate image recognition and classification.

3.2.2 Threshold setting and performance analysis

To achieve ITSM for assisting construction site safety inspections, this study developed a text description database based on relevant safety regulations. This database serves as the semantic basis for the CLIP model, containing descriptions of various non-compliance scenarios related to construction opening protection measures. These descriptions cover

common safety violations, such as "excessive guardrail spacing" and "incorrect guardrail installation". During system operation, the CLIP model processes input images and selects the most appropriate violation description from the database to determine whether a given image corresponds to a specific non-compliance scenario.

To ensure the accuracy and reliability of the ITSM process, this study conducted threshold optimization and sensitivity analysis for the CLIP model. The threshold value directly influences the sensitivity and accuracy of the violation detection system:

Excessively high threshold: If the threshold is set too high, the CLIP model will only classify an image as a violation if its similarity score exceeds the defined threshold. This could result in **false negatives**, where certain non-compliant scenarios remain undetected due to lower similarity scores. Consequently, some potential hazards may be overlooked by on-site personnel, compromising the comprehensiveness of safety management.

Excessively low threshold: If the threshold is set too low, even images with relatively low similarity scores may be classified as violations, leading to **false positives**. This would generate an excessive number of reports containing non-violative conditions, thereby increasing the workload of safety inspectors and reducing the practical utility of the system.

The image-text matching process is illustrated in Figure 2. The system first captures images of the construction site, which are processed through an image encoder to extract visual feature vectors. Simultaneously, predefined textual descriptions of potential safety violations (e.g., "the spacing between guardrails is too wide") are input into a text encoder to generate corresponding textual feature vectors.

After feature extraction, visual and textual features are projected into a shared embedding space. Within this space, the system performs semantic matching by calculating the cosine similarity between the feature vectors. The similarity score represents the semantic alignment between the site image and the violation descriptions. In the example shown in Figure 2, the calculated cosine similarity is 0.72. If this score exceeds the system's predefined threshold, the system automatically classifies the image as a "violation."

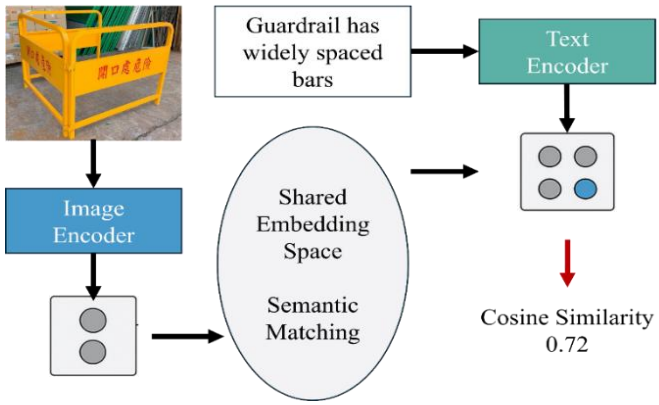

Figure 2. Semantic matching process for construction site safety compliance using CLIP

3.3 System framework and workflow

Figure 3 illustrates the complete system architecture for the automated construction safety inspection process, which is divided into four main stages: Data Input, Object Detection, Semantic Matching, and Risk Report Generation, achieving real-time monitoring and risk management.

First, the system starts with the Input Phase, where users upload construction site images as the foundational data for subsequent analysis. These images undergo Standardize Image processing to ensure consistent size and resolution, enhancing the accuracy and stability of the model analysis. After standardization, the images enter the Detect Image process, where the YOLO model automatically detects construction openings, guardrails, fences, and fall protection installations, marking potential hazard areas for further evaluation.

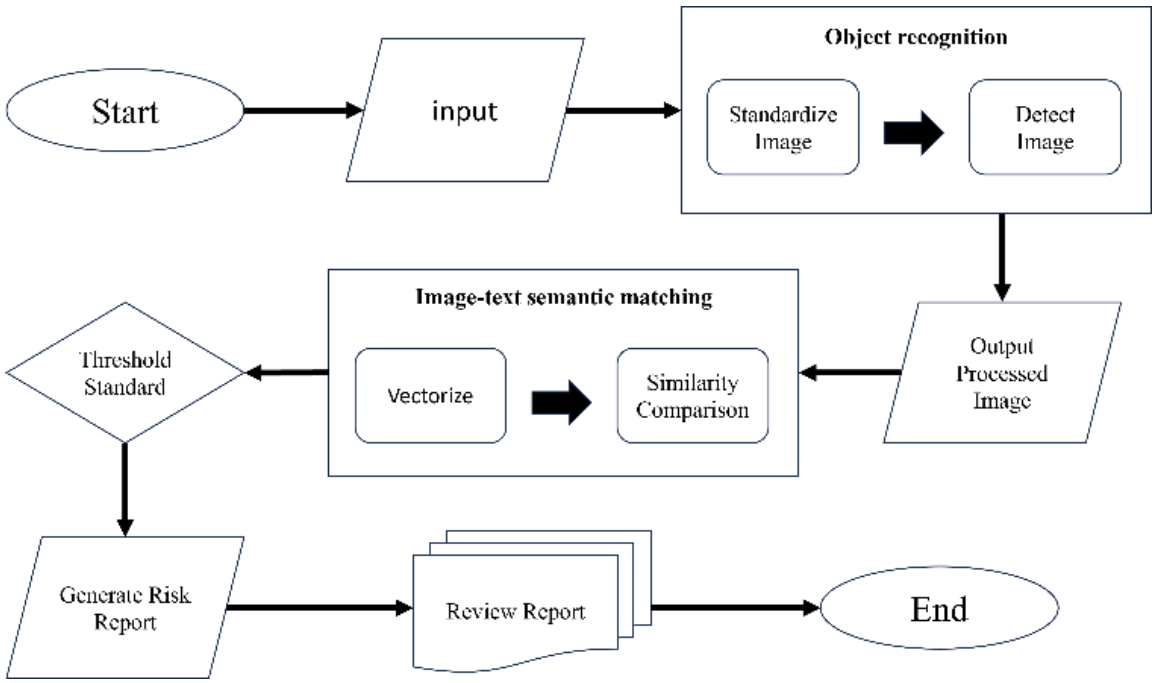

Figure 3. System framework

Next, as shown in Figure 3, the detected images proceed to the Vectorize stage, where image features are encoded into high-dimensional vectors and compared with predefined safety regulations. This process utilizes Similarity Comparison to calculate the cosine similarity between image and text descriptions, determining whether the site installations comply with safety standards. If the similarity score exceeds the system-defined Threshold Standard, the image is flagged as a "Potential Violation" for further processing.

Following the similarity matching, the system automatically enters the Generate Risk Report stage, generating structured reports for the identified potential hazards. These reports detail the location, type, and risk level of the non-compliant installations. The generated reports are then presented in the Review Report interface, allowing site management personnel to review and handle the detected issues, enhancing the transparency and traceability of inspection results.

Finally, after the inspection and report generation processes are complete, the system transitions to the End phase, indicating the completion of that batch of inspections. As depicted in Figure 3, this entire workflow achieves full automation from data collection to semantic matching, risk identification, report generation, and review. It not only reduces the costs associated with manual inspections but also enables real-time detection of potential construction hazards, ensuring site safety and improving management efficiency.

4. RESULTS

This section discusses the results of the Object Recognition model and image-text semantic matching. In the Object Recognition part, the analysis focuses on the model's accuracy and stability in detecting construction openings and fall protection facilities. In the image-text semantic matching part, the effectiveness of the CLIP model in comparing images with safety regulations is examined, along with its recognition capabilities under zero-shot learning to ensure compliance with safety standards.

4.1 Object Recognition model results

Table 4 compares the performance of various YOLO models (YOLOv8 and YOLOv10 series) in construction site image recognition, including Precision, Recall, mAP@50, mAP@50-95, and Training Time. The results indicate that model performance improves as the number of parameters increases. Among the tested models, YOLOv8l demonstrated outstanding performance across all key metrics, particularly in terms of precision, recall, and mAP@50, showing high detection accuracy and stability. Additionally, YOLOv8l required a relatively shorter training time. Therefore, YOLOv8l was ultimately selected as the object detection model for the proposed system.

To further analyze the performance differences among various YOLO versions, this study summarizes the number of layers, parameters, and computational complexity (GFLOPs) for each model, as shown in Table 5. Overall, the results indicate a positive correlation between the number of parameters and detection performance metrics (mAP, precision, and recall). However, increasing the number of parameters also significantly extends training time. For instance, although YOLOv10x offers high accuracy, its training time reaches up to 67 hours, making it less ideal for construction sites that require rapid deployment.

Table 4. Object detection model data comparison

Model	Precis -ion	Recall	mAP @50	mAP @50-95	Training Time (hr)
YOLOv8n	0.815	0.823	0.802	0.764	1.90
YOLOv8s	0.861	0.876	0.853	0.821	1.98
YOLOv8m	0.902	0.907	0.892	0.858	3.88
YOLOv8l	0.943	0.952	0.923	0.914	5.78
YOLOv8x	0.946	0.923	0.925	0.915	9.46
YOLOv10n	0.785	0.803	0.791	0.732	1.84
YOLOv10s	0.814	0.825	0.805	0.774	2.85
YOLOv10m	0.865	0.874	0.859	0.823	4.20
YOLOv10b	0.902	0.894	0.877	0.852	5.33
YOLOv10l	0.920	0.912	0.910	0.885	5.20
YOLOv10x	0.925	0.915	0.914	0.886	67.00

Table 5. Model summary

Versions	Layers	Parameters	GFLOPs
YOLOv8n	218	3,006,428	8.1
YOLOv8s	168	11,127,132	28.4
YOLOv8m	218	25,842,076	78.7
YOLOv8l	268	43,609,692	164.68
YOLOv8x	268	68,127,420	257.4
YOLOv10n	102	2,265,948	6.5
YOLOv10s	106	7,219,548	21.4
YOLOv10m	136	15,315,484	58.9
YOLOv10b	142	19,007,196	91.6
YOLOv10l	174	24,312,412	120.0
YOLOv10x	192	29,400,380	160.0

YOLOv8 and YOLOv10 both demonstrated robust detection accuracy in this study. YOLOv8 adopts a decoupled head and lightweight backbone, which shortens training time and supports rapid deployment on medium-scale datasets. In contrast, YOLOv10 incorporates Unified Label Assignment and Dynamic Head mechanisms, which improve detection for small-scale objects, but require longer training time due to its deeper network and higher parameter complexity.

Table 6. Comparison between YOLOv8 and YOLOv10

Item	YOLOv8	YOLOv10
Execution Speed	Faster inference and training	Slightly slower but more stable
Detection of Large Safety Structures	High accuracy	Comparable or slightly higher
Detection of Small or Occluded Objects	Moderate performance	Better capability in detecting small objects
Model Size and Deployment Feasibility	Lightweight model, suitable for edge deployment	Larger model, suitable for resource-rich environments

As shown in Table 6, both YOLOv8 and YOLOv10 performed well in detecting fall protection facilities around construction openings. Among them, YOLOv8l yielded high precision, recall, and mAP scores, while maintaining shorter inference times, making it well-suited for on-site inspection systems that demand rapid feedback. In particular, YOLOv8 demonstrated faster inference and training speeds, improved detection accuracy for large safety structures such as guardrails, and a lightweight architecture conducive to edge

deployment.

On the other hand, YOLOv10x achieved slightly higher mAP scores and showed superior performance in detecting small or occluded objects. However, its training time reached 67 hours, and due to its higher model complexity, it is more appropriate for deployment in resource-rich environments rather than in real-time field applications.

4.2 Image-text semantic matching results

Table 7 presents the impact of different threshold settings on the detection accuracy of the proposed system. The results show that when the threshold is set to 0.66, the system achieves the highest detection accuracy of 85% (28/33), indicating that this threshold effectively balances FP and FN.

Table 7. Detection performance by threshold level

Threshold	Accuracy
0	18/33=55%
0.65	26/33=79%
0.66	28/33=85%
0.67	22/33=67%
0.68	15/33=45%
0.69	3/33=9%
0.70	3/33=9%
0.75	0/33=0%

At lower thresholds (e.g., 0 or 0.65), the system can still detect some violation scenarios, but with a higher risk of false positives. However, the detection accuracy drops significantly as the threshold gradually increases to 0.67 or above. In particular, when the threshold reaches 0.75, the system is entirely unable to detect any violations, demonstrating that excessively high thresholds overly restrict the matching results between images and text.

This observation reflects that threshold settings have a decisive impact on detection performance in the image-text semantic matching process of the CLIP model. If the threshold is too high, potential risks may go undetected; if it is too low, it may cause a surge in false positives. Therefore, this study selects 0.66 as the optimal threshold to achieve a balance between detection accuracy and stability.

4.3 System interface

Figure 4 illustrates the complete interface workflow of the automated detection and report generation system developed in this study. Upon entering the system, users first access the basic information input interface, as shown in Figure 4(a). At this stage, users are required to fill in basic information such as project name, person in charge, inspection date, and inspection location to ensure that the subsequent analysis report can accurately correspond to the specific location and time of the construction site. After completing the information input, users click the "Run" button, and the system automatically performs image recognition and image-text matching, leading to the directory list page as depicted in Figure 4(b).

In Figure 4(b), the system displays all previously analyzed records in a directory format, allowing users to search for historical data quickly. Each directory item is labeled with date and location information. Users only need to click on the corresponding directory to access the detailed report view page, as shown in Figure 4(c). The system automatically organizes all detected deficiencies from the analysis into a structured report on this interface.

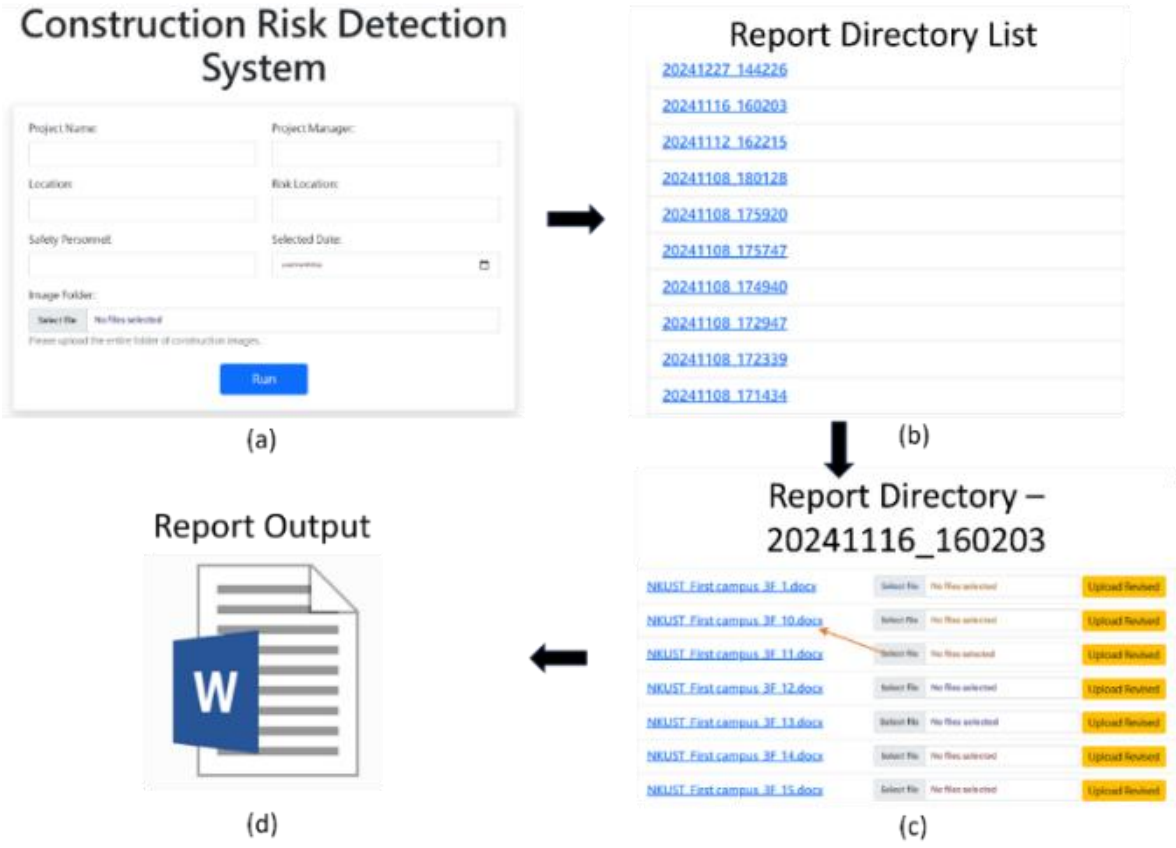

Figure 4. System interface display

Opening Risk Safety Advice Report

project name	NKUST	Project leader	George
Place	First campus	record date	2024-11-16
risk location	3F	unsafe condition	The steel pipe guardrail lacks one of the upper railing, middle railing, and pillar.
Detailed description	Insufficient guardrail components (steel pipes)	Safety Officer	David

Picture before improvement

Reference

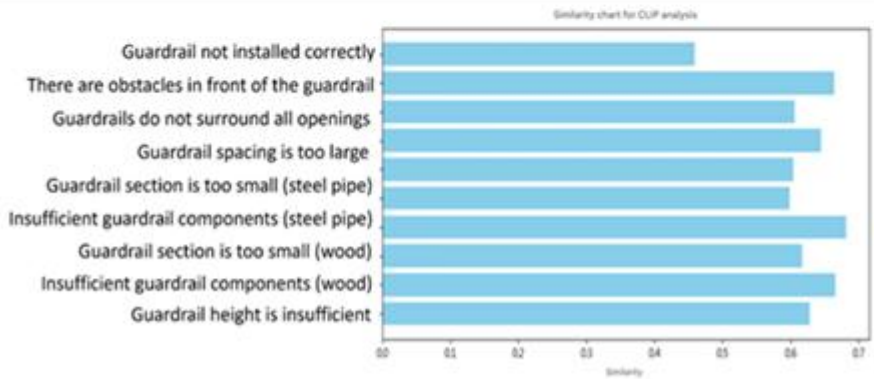

Figure 5. Report display

In Figure 4(c), users can select any report file, and the system will display detailed reasons for violations, image evidence, and corresponding standard regulations, helping management personnel quickly understand the issues and take corrective actions. Additionally, if on-site corrections and improvements have been made, the system provides an update upload function, allowing users to instantly upload the latest image records and perform a secondary inspection through the system to verify whether the corrections meet safety standards. This design accelerates the handling efficiency of on-site deficiencies and ensures transparency and real-time updates of inspection results, effectively enhancing the safety management level of construction sites.

The final report format is shown in Figure 5, which provides a detailed presentation of the basic information entered by the user through the interface, along with the system's analysis results. The report includes the original data provided by the user and the conclusions derived from the analysis tools. It offers a comprehensive risk assessment and recommendations, enabling relevant personnel to understand the on-site situation and take appropriate measures.

5. CONCLUSION AND RECOMMENDATIONS

In traditional construction site management, inspecting fall protection facilities around openings primarily relies on manual visual inspection. This method is time-consuming and prone to errors due to subjective judgment, making it challenging to ensure that all openings meet safety standards. Particularly in large-scale construction projects, the numerous and widely distributed openings increase inspection difficulty and labor costs, resulting in many potential safety risks not being identified in time, thereby raising the risk of falls. According to statistics from Taiwan's Occupational Safety and Health Administration, falls from heights are among the most fatal accidents in the construction industry, accounting for a major proportion of construction site incidents.

To address these issues, this study developed an automated detection and report generation system for construction site opening safety risks based on image recognition technology. The system integrates the YOLO model for object detection, enabling rapid identification of openings and the presence of fall protection facilities in construction site images. Simultaneously, it employs the CLIP model for image-text semantic matching, effectively filtering image content that aligns most closely with violation descriptions and automatically generating structured risk reports. The research results indicate that the system achieves high accuracy in identifying unprotected openings and reaches an accuracy rate of 85% in detecting non-compliant protective installations, demonstrating significant detection performance.

To further enhance the system's accuracy and coverage in detecting fall protection facilities around construction site openings, future research is suggested to focus on two main directions. First, expanding the image dataset and diversifying the types of protective facilities, including more varied construction scenarios and different materials and forms of protective equipment, such as safety lines, temporary guardrails, and more. By increasing the diversity of image samples, the deep learning model's recognition capability in different construction environments can be improved, enhancing the system's ability to assess the compliance of protective facilities. Furthermore, richer image data can help the model learn better about various site layouts and complex backgrounds of fall protection facilities around openings, reducing false positives and false negatives, and improving overall detection efficiency.

Secondly, although this study focuses on the automated detection and report generation of fall protection facilities around openings, risk management at construction sites is not limited to fall protection. It also includes various safety risks such as electric shocks, slips, falling objects, and heavy equipment entrapment. Therefore, future research should consider integrating detection technologies for other hazards into the system architecture, building a more comprehensive construction safety monitoring system. By integrating multiple hazard types, the system can instantly detect potential dangers at construction sites and automatically generate diversified risk detection reports, helping management personnel accurately and comprehensively understand the site's safety status. This multi-dimensional hazard recognition model is expected to significantly strengthen disaster prevention capabilities at construction sites, improve the accuracy and response speed of accident prevention, achieve higher safety management standards, and promote the digital transformation of construction safety management.

REFERENCES

[1] Schneider, S., Susi, P. (1994). Ergonomics and construction: A review of potential hazards in new construction. American Industrial Hygiene Association Journal, 55(7): 635-649. https://doi.org/10.1080/15428119491018727

[2] Nadhim, E.A., Hon, C., Xia, B., Stewart, I., Fang, D. (2016). Falls from height in the construction industry: A critical review of the scientific literature. International Journal of Environmental Research and Public Health, 13(7): 638. https://doi.org/10.3390/ijerph13070638

[3] Lombardi, R., Secundo, G. (2021). The digital transformation of corporate reporting – A systematic literature review and avenues for future research. Meditari Accountancy Research, 29(5): 1179-1208. https://doi.org/10.1108/MEDAR-04-2020-0870

[4] Wong, L., Wang, Y., Law, T., Lo, C.T. (2016). Association of root causes in fatal fall-from-height construction accidents in Hong Kong. Journal of Construction Engineering and Management, 142(7): 04016018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001098

[5] Akcay, E.C., Arditi, D. (2022). Predicting employer and worker responsibilities in accidents that involve falls in building construction sites. Buildings, 12(4): 464. https://doi.org/10.3390/buildings12040464

[6] Kelm, A., Laußat, L., Meins-Becker, A., Platz, D., Khazaee, M.J., Costin, A.M., Helmus, M., Teizer, J. (2013). Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. Automation in Construction, 36: 38-52. https://doi.org/10.1016/j.autcon.2013.08.009

[7] Shrestha, K., Shrestha, P.P., Bajracharya, D., Yfantis, E.A. (2015). Hard-hat detection for construction safety visualization. Journal of Construction Engineering, 2015(1): 721380. https://doi.org/10.1155/2015/721380

[8] Zhang, H., Yan, X., Li, H., Jin, R., Fu, H. (2019). Real-

time alarming, monitoring, and locating for non-hard-hat use in construction. Journal of Construction Engineering and Management, 145(3): 04019006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001629
[9] Zhu, Z., Sun, X., Zhao, F., Meixner, F.X. (2015). Ozone concentrations, flux and potential effect on yield during wheat growth in the Northwest-Shandong Plain of China. Journal of Environmental Sciences, 34: 1-9. https://doi.org/10.1016/j.jes.2014.12.022
[10] Fang, W., Love, P.E.D., Ding, L., Xu, S., Kong, T., Li, H. (2023). Computer vision and deep learning to manage safety in construction: Matching images of unsafe behavior and semantic rules. IEEE Transactions on Engineering Management, 70(12): 4120-4132. https://doi.org/10.1109/TEM.2021.3093166
[11] Ding, Y., Liu, M., Luo, X. (2022). Safety compliance checking of construction behaviors using visual question answering. Automation in Construction, 144: 104580. https://doi.org/10.1016/j.autcon.2022.104580
[12] Tang, W. (2023). Application of BIM technology in the reinforcement and renovation of existing building inspection projects. Alexandria Engineering Journal, 82: 240-247. https://doi.org/10.1016/j.aej.2023.09.075
[13] Goh, Y.M., Ubeynarayana, C.U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. Accident Analysis & Prevention, 108: 122-130. https://doi.org/10.1016/j.aap.2017.08.026
[14] Hartmann, T., Trappey, A. (2020). Advanced engineering informatics—Philosophical and methodological foundations with examples from civil and construction engineering. Developments in the Built Environment, 4: 100020. https://doi.org/10.1016/j.dibe.2020.100020
[15] Boedker, C., Mouritsen, J., Guthrie, J. (2008). Enhanced business reporting: International trends and possible policy directions. Journal of Human Resource Costing & Accounting, 12(1): 14-25. https://doi.org/10.1108/14013380810872734
[16] Alles, M.G., Dai, J., Vasarhelyi, M.A. (2021). Reporting 4.0: Business reporting for the age of mass customization. Journal of Emerging Technologies in Accounting, 18(1): 1-15. https://doi.org/10.2308/jeta-10764
[17] Zhou, Z., Goh, Y.M., Li, Q. (2015). Overview and analysis of safety management studies in the construction industry. Safety Science, 72: 337-350. https://doi.org/10.1016/j.ssci.2014.10.006
[18] Birhane, G. E., Yang, L., Geng, J., Zhu, J. (2022). Causes of construction injuries: A review. International Journal of Occupational Safety and Ergonomics, 28(1): 343-353. https://doi.org/10.1080/10803548.2020.1761678
[19] Kayastha, R., Kisi, K. (2024). Assessing factors affecting fall accidents among Hispanic construction workers: Integrating safety insights into BIM for enhanced life cycle management. Buildings, 14(9): 3017. https://doi.org/10.3390/buildings14093017
[20] Chan, A., Yang, Y., Choi, T., Nwaogu, J. (2022). Characteristics and causes of construction accidents in a large-scale development project. Sustainability, 14(8): 4449. https://doi.org/10.3390/su14084449
[21] Rabbi, A.B.K., Jeelani, I. (2024). AI integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications. Automation in Construction, 164: 105443. https://doi.org/10.1016/j.autcon.2024.105443
[22] Guo, B., Zou, Y., Chen, L. (2018). A review of the applications of computer vision to construction health and safety. University of Auckland: Auckland, New Zealand.
[23] C, V., Salve, U.R. (2020). A scientometric analysis and review of fall from height research in construction. Construction Economics and Building, 20(1): 17-35. https://doi.org/10.5130/AJCEB.v20i1.6802
[24] Laad, M., Maurya, R., Saiyed, N. (2024). Unveiling the vision: A comprehensive review of computer vision in AI and ML. In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, pp. 1-6. https://doi.org/10.1109/ADICS58448.2024.10533631
[25] Sinha, R.K., Pandey, R., Pattnaik, R. (2018). Deep learning for computer vision tasks: A review. arXiv preprint arXiv:1804.03928. https://doi.org/10.48550/arXiv.1804.03928
[26] Terven, J., Cordova-Esparza, D. (2024). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. arXiv preprint arXiv:2304.00501. https://doi.org/10.48550/arXiv.2304.00501
[27] Zhang, Z., Li, H., Guo, H., Wu, Y., Luo, Z. (2024). Causal inference of construction safety management measures towards workers' safety behaviors: A multidimensional perspective. Safety Science, 172: 106432. https://doi.org/10.1016/j.ssci.2024.106432
[28] Kim, S., Hong, S.H., Kim, H., Lee, M., Hwang, S. (2023). Small object detection (SOD) system for comprehensive construction site safety monitoring. Automation in Construction, 156: 105103. https://doi.org/10.1016/j.autcon.2023.105103
[29] Wang, Z., Wu, Y., Yang, L., Thirunavukarasu, A., Evison, C., Zhao, Y. (2021). Fast personal protective equipment detection for real construction sites using deep learning approaches. Sensors, 21(10): 3478. https://doi.org/10.3390/s21103478
[30] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458. https://doi.org/10.48550/arXiv.2405.14458
[31] Lahajal, N.K., S, H. (2024). Enhancing image retrieval: A comprehensive study on photo search using the CLIP mode. arXiv preprint arXiv:2401.13613. https://doi.org/10.48550/arXiv.2401.13613
[32] Lülf, C., Lima Martins, D.M., Vaz Salles, M.A., Zhou, Y., Gieseke, F. (2024). CLIP-branches: Interactive fine-tuning for text-image retrieval. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, pp. 2719-2723. https://doi.org/10.1145/3626772.3657678
[33] Fu, Z., Mao, Z., Song, Y., Zhang, Y. (2023). Learning semantic relationship among instances for image-text matching. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 15159-15168. https://doi.org/10.1109/CVPR52729.2023.01455
[34] Cheng, R., Cui, W. (2024). Image-text matching with multi-view attention. arXiv preprint arXiv:2402.17237. https://doi.org/10.48550/arXiv.2402.17237
[35] Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R. (2022). The unreasonable effectiveness of

CLIP features for image captioning: An experimental analysis. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp. 4661-4669. https://doi.org/10.1109/CVPRW56347.2022.00512

[36] Che, C., Lin, Q., Zhao, X., Huang, J., Yu, L. (2023). Enhancing multimodal understanding with CLIP-based image-to-text transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies, Qingdao, China, pp. 414-418. https://doi.org/10.1145/3627377.3627442

[37] Liu, X., Du, R., Tan, L., Xu, J., Chen, C., Jiang, H., Aldwais, S. (2024). CIB-SE-YOLOv8: Optimized YOLOv8 for real-time safety equipment detection on construction sites (Version 1). arXiv preprint arXiv:2410.20699. https://doi.org/10.48550/ARXIV.2410.20699

[38] Zhang, Y., Guan, D., Zhang, S., Su, J., Han, Y., Liu, J. (2024). GSO-YOLO: Global stability optimization YOLO for construction site detection. arXiv preprint arXiv:2407.00906. https://doi.org/10.48550/arXiv.2407.00906

[39] Ding, Y., Luo, X. (2023). Personal protective equipment detection in extreme construction conditions (Version 1). arXiv preprint arXiv:2307.13654. https://doi.org/10.48550/ARXIV.2307.13654

[40] Tsai, W.L., Le, P.L., Ho, W.F., Chi, N.W., Lin, J.J., Tang, S., Hsieh, S.H. (2025). Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention. Automation in Construction, 169: 105863. https://doi.org/10.1016/j.autcon.2024.105863

[41] Zuguo, C., Aowei, K., Yi, H., Jie, J. (2025). Enhanced PEC-YOLO for detecting improper safety gear wearing among power line workers (Version 1). arXiv preprint arXiv:2501.13981. https://doi.org/10.48550/ARXIV.2501.13981

[42] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020

[43] Theisen, W., Scheirer, W. (2023). C-CLIP: Contrastive image-text encoders to close the descriptive-commentative gap. arXiv preprint arXiv:2309.03921. https://doi.org/10.48550/arXiv.2309.03921

[44] Dumouchelle, J., Frejinger, E., Lodi, A. (2023). Reinforcement learning for freight booking control problems. arXiv preprint arXiv:2102.00092. https://doi.org/10.48550/arXiv.2102.00092