EXPLAINABLE AI-DRIVEN UAVS FOR REAL-TIME FALL-FROM-HEIGHT RISK MONITORING IN CONSTRUCTION

KONG TING

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Building and Real Estate

Explainable AI-Driven UAVs for Real-Time Fall-From-Height Risk Monitoring in Construction

KONG Ting

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

_____(Signed)

KONG Ting (Name of student)

ABSTRACT

Falls from height (FFH) have been consistently identified as a leading cause of injuries and fatalities in the construction industry, with most fall-related incidents attributable to the absence of or inadequate safety protections (e.g., unprotected open edges). Typically, traditional approaches heavily rely on manual inspections, which pose numerous challenges (i.e., being time-consuming) in real-life construction sites due to the nature of their complex and dynamic working conditions.

To overcome these shortcomings, emerging technologies such as Building Information Modelling (BIM), Deep Learning (DL), and Computer Vision (CV) offer promising solutions for the automatic detection of hazards to improve safety performance in construction. Despite their success, the following limitations hinder their deployment and applications on construction sites: (1) existing deep learning models are complex non-linear statistical models, which suffer significantly from the "black box" problem, making it difficult for on-site managers to trust the outcomes produced by such AI systems; (2) prior computer vision approaches primarily rely on 2D/3D images, which often cannot localize hazards precisely within 3D models or associate the hazards with specific building components; and (3) many studies tend to be retrospective and focus on examining the conditions that contribute to unsafe behaviour from a psychological perspective, which are unable to visually comprehend the dynamic and complex conditions that influence unsafe behaviour.

To address these limitations, our research aims to investigate the following research question: *How could we improve the accuracy and interpretability of deep learning models in monitoring Fall-From-Height (FFH) risk in construction sites by using Unmanned Aerial Vehicles (UAVs) and computer vision systems*? To this end, we adopted a design science research (DSR) paradigm to guide the conceptual framework design, model development, and evaluation. Following such a design science approach, we design, implement, and evaluate an XAI-enabled UAV-based computer vision system that (1) detects unprotected edges and guardrails, (2) precisely localizes hazards related to FFH within a three-dimensional (3D) construction site model and links the hazards to specific building components, and (3) predicts workers approaching unprotected edges. Real-life projects are used to validate the effectiveness and feasibility of the proposed approach.

The research findings presented in this research have demonstrated that the proposed explainable Artificial Intelligence (XAI)-based computer vision system is able to accurately identify hazards. We suggest that our research can be used to improve worker safety and reduce operational costs by replacing labour-intensive inspections with automated monitoring, delivering a transparent, resilient, and practical safety solution for workers, safety managers, regulators, and the construction industry.

PUBLICATIONS ARISING FROM THE THESIS

*An asterisk * indicates the corresponding author. Some chapters of the thesis are based on parts of the published papers, which have been noted in the corresponding contents.*

Journal Papers (Published)

1. **Kong, T.**, Fang, W., Love, P. E.D, Luo, H., Xu, S., & Li, H. (2021). Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction. *Advanced Engineering Informatics*, 50, 101400.
2. Fang, X., Li, H., Wu, H., Fan, L., **Kong, T.***, & Wu, Y. (2023). A fast end-to-end method for automatic interior progress evaluation using panoramic images. *Engineering Applications of Artificial Intelligence*, 126, 106733.
3. Li, M., Ren, Q., Li, M., **Kong, T**., Li, H., Tian, H., & Liu, S. (2024). Digital twin-enabled collision early warning system for marine piling: Application to a wharf project in China. *Advanced Engineering Informatics*, 59, 102269.
4. Ren, Q., Li, H., Li, M., **Kong, T**., & Guo, R. (2023). Bayesian incremental learning paradigm for online monitoring of dam behavior considering global uncertainty. *Applied Soft Computing*, 143, 110411.
5. Fang, W., Love, P. E.D., Ding, L., Xu, S., **Kong, T.**, & Li, H. (2021). 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.
6. Ren, Q., Li, H., Li, M., Zhang, J., & **Kong, T**. (2023). Towards online monitoring of concrete dam displacement subject to time-varying environments: An improved sequential learning approach. *Advanced Engineering Informatics*, 55, 101881.
7. Ren, Q., Li, H., Zheng, X., Li, M., Xiao, L., & **Kong, T**. (2023). Multi-block synchronous prediction of concrete dam displacements using MIMO machine learning paradigm. *Advanced Engineering Informatics*, 55, 101855.
8. Ren, Q., Li, M., **Kong, T.**, & Ma, J. (2022). Multi-sensor real-time monitoring of dam behavior using self-adaptive online sequential learning. *Automation in Construction*,

140, 104365.

9. Zhou, C., **Kong, T**., Jiang, S., Chen, S., Zhou, Y., & Ding, L. (2020). Quantifying the evolution of settlement risk for surrounding environments in underground construction via complex network analysis. *Tunnelling and Underground Space Technology*, 103, 103490.
10. Zhou, C., **Kong, T.**, Zhou, Y., Zhang, H., & Ding, L. (2019). Unsupervised spectral clustering for shield tunneling machine monitoring data with complex network theory. *Automation in Construction*, 107, 102924.

ACKNOWLEDGEMENTS

At the close of my PhD, I would like to express my heartfelt gratitude to everyone who supported and encouraged me through the many ups and downs of these years. I had imagined countless times how I would summarise this academic journey and write these acknowledgements, yet I often found myself lost at the final step. Fortunately, with the perseverance of those around me-and my own-I have reached this moment. This unique experience has helped me grow and strengthen my belief in the power of perseverance through both laughter and tears.

First and foremost, I would like to express my sincere gratitude to my chief supervisor, Professor Heng LI. He gave me the opportunity and unwavering support to pursue my PhD dream. His passion for research, valuable experience, and deep insight continually inspired me to explore new ideas within and across disciplines. From refining the research direction, defining the research gap, and completing the thesis, Prof. Li was always available to provide support. I am especially grateful for his patience and for allowing me to progress at my own pace. Prof. Li is an extraordinary mentor and person; he has taught me to be self-motivated and to share ideas actively. It is an invaluable privilege to have had such a remarkable supervisor.

I would also like to express my special thanks to Dr Xiaoying LI, Dr Xin FANG, Dr Qiubin REN, Dr Haitao WU, Dr Jiangjiang YU, Dr Yi YU, Ms Yue WU, and all colleagues from

the Smart Construction Laboratory and the Department of Building and Real Estate for their support and care-both academically and personally-which made my time in Hong Kong unforgettable. It has been an honour to meet and work with such excellent team members. My gratitude also extends to the experts and scholars who examined this thesis for their careful reviews and constructive suggestions.

Finally, I owe my Loving thanks to my parents. Thank you for standing by my side in every moment of my life. You mean everything to me, and I will always strive for our family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ILO International Labour Organization

BLS Bureau of Labor Statistics

MOHURD China's Ministry of Housing and Urban-Rural Development

TRA Theory of Reasoned Action

TPB Theory of Planned Behaviour

ANN Artificial neural networks

IoT Internet of Things

YOLO You Only Look Once

BIM Building Information Modeling

AI Artificial Intelligence

UAV Unmanned Aerial Vehicles

XAI Explainable Artificial Intelligence

FFH Fall from Heights

CACS Construction Accident Causality System

AR Augmented Reality

CV Computer Vision

CNN Convolutional Neural Network

RNN Recurrent Neural Network

LSTM Long Short-Term Memory

PPE Personal Protective Equipment

SSD Single Shot-multi-box Detector

VOT Visual object tracking

SiamFC Siamese framework

CAM Class Activation Mapping

LIME Local Interpretable Model-agnostic Explanations

RISE Randomized Input Sampling for Explanation

SHAP SHapley Additive exPlanations

LRP Relevance Propagation

TCAV Concept Activation Vectors

FRFN Feature Refinement Feed-Forward Network

SAFM Spatial Adaptive Feature Modulation

C2PSA Cross-Stage Partial with Pyramid Squeeze Attention

DWConv Depthwise Convolution

GAP Global Average Pooling

TP True Positive

FP False Positive

FN False Negatives

CBAM Convolutional Block Attention Module

AP Average Precision

SE Squeeze-and-Excitation

IoU Intersection over Union

CHAPTER 1 INTRODUCTION[1]

This chapter commences by presenting the workforce safety issue, specifically, falls from height (FFH) risk in the construction sector, and emphasizes the significance of examining edge protection for the improvement of safety performance. To provide context for the research topic, this chapter gives an overview of existing technologies that have been used to detect hazards related to falls from height. Next, we define the research aims and objectives of this thesis, outline its structure, and highlight its research significance.

1.1 Background

The global construction industry has expanded rapidly in recent years with its profound impact on national economies. Nevertheless, high levels of safety risk in construction have been identified as one of its most persistent challenges (Stemn *et al*., 2018; Kulinan *et al*., 2024). Construction sites are inherently complex, dynamic, and labor-intensive environments where on-site participants are easily exposed to various hazards. Notably, frequent accidents on construction sites often have devastating consequences, including delays in project schedules, significant economic losses, or even endangering workers' lives, which can trigger a series of social problems (Man *et al*., 2021; Lafhaj *et al*., 2024; Kumi *et al*., 2024).

[1] This chapter is based on a published study and being reproduced with the permission of Elsevier.

Kong, T., Fang, W., Love, P. E.D, Luo, H., Xu, S., & Li, H. (2021). Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction. Advanced Engineering Informatics, 50, 101400.

According to the International Labour Organization (ILO), more than 30% of industrial fatalities worldwide are attributed to the construction sector, which holds the highest proportion among all industries in 2022 (ILO, 2022). Although construction workers represent only 5% of the workforce in the United Kingdom, the sector accounts for 27% of all workplace fatalities (HSE, 2024). In the United States (U.S.), 1,092 workers lost their lives while being engaged in construction activities in 2022, marking a continuation of a five-year upward trend (BLS, 2022). The U.S. Bureau of Labor Statistics (BLS) also reported that between 2016 and 2022, annual occupational injuries and illnesses related to the construction industry exceeded over sixty thousand on average. Similarly, in Hong Kong, the Development Bureau (2023) reported 3,097 worksite accidents in 2023, confirming construction as one of the most hazardous industries in the region (HK Labour Department, 2023). Occupational injuries and illnesses, construction-related fatal injuries in the U.S., as well as industrial accidents in Hong Kong's construction industry are presented in Figures 1-3, respectively. These statistics underscore the urgent need for developing effective and proactive safety measures to prevent accidents on construction sites.

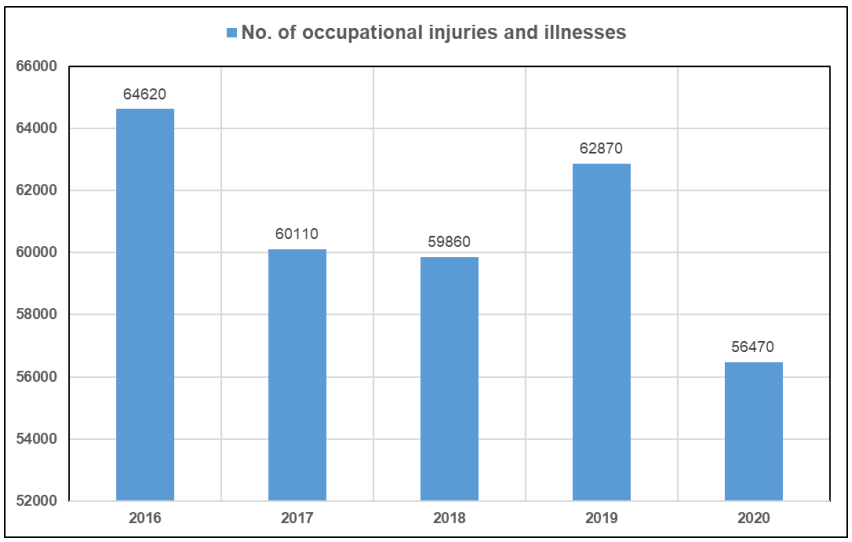


Figure 1 Occupational Injuries and Illnesses in Construction Industry of U.S. (2016-2020)

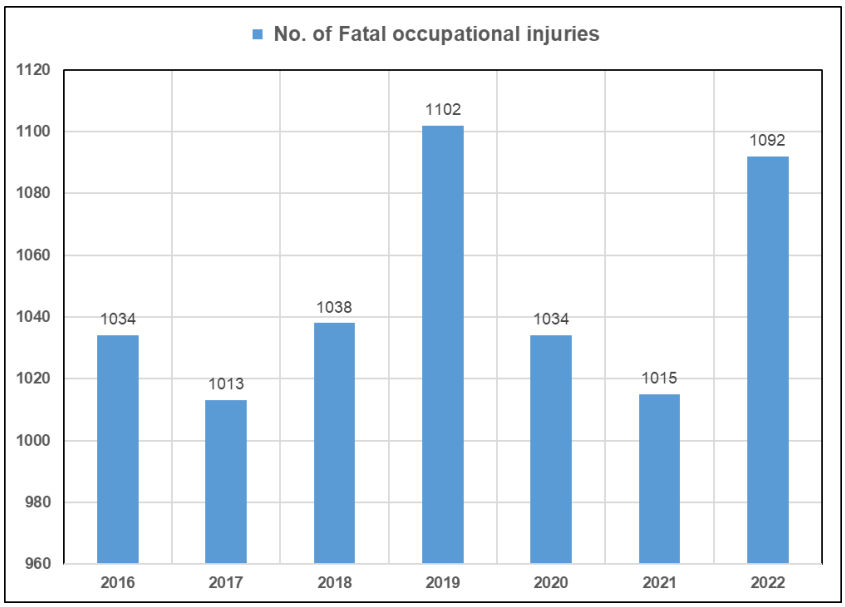


Figure 2 Fatal Occupational Injuries in Construction Industry of U.S. (2016-2022)

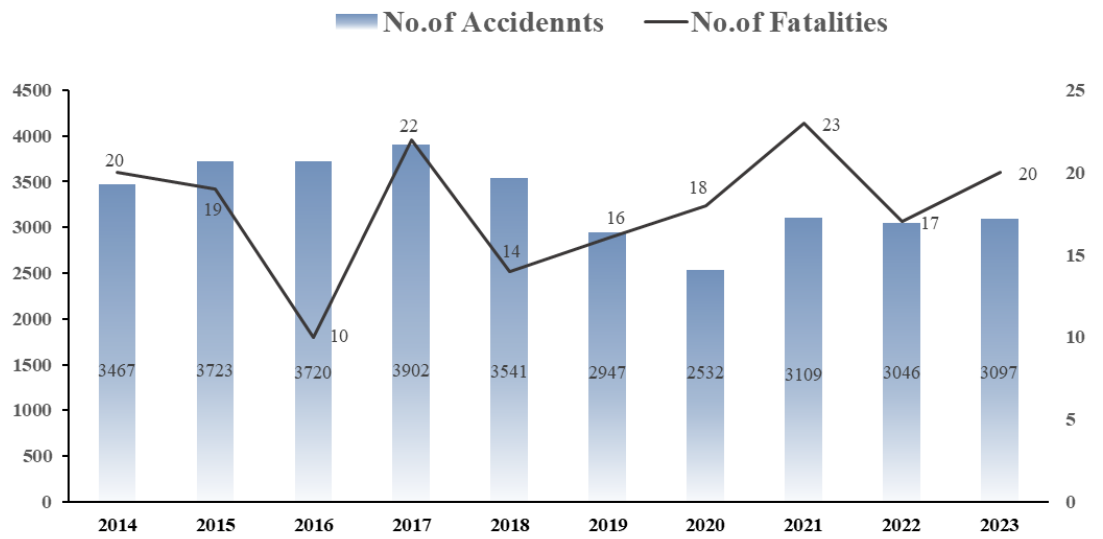


Figure 3 Industrial Accidents in Construction Industry of Hong Kong (2015-2024)

Falls from height (FFH) are consistently the leading cause of fatalities in construction, while many workers can hardly avoid working at height due to daily tasks (Kang *et al*., 2022; Qi *et al*., 2025). In 2020, the China Ministry of Housing and Urban-Rural Development (MOHURD, 2022) reported that FFH accounted for 59.07% of all construction accidents, followed by being struck by objects (12.05%), as presented in Figure 4. Similarly, statistics from the Hong Kong Labour Department in 2023 also demonstrate that falls rank among the Top3 accident types in the construction industry, as noted in Figure 5 (HK Labour Department, 2023).

Researchers have noted that most fall-related incidents are caused by missing or inadequate safety protections, such as unprotected open edges/skylights, slippery surfaces, and improperly positioned ladders (Ding *et al*., 2018; Li *et al.*, 2019; Oliveira *et al*., 2023). Additionally, some fall protection systems may be removed during work activities in the ever-changing construction environments (Chain *et al*., 2021). For

example, on a high-rise project in Shenzhen, China, three workers fell from the 42nd floor after temporary guardrails were removed but not reinstalled, resulting in two fatalities and one critical injury.

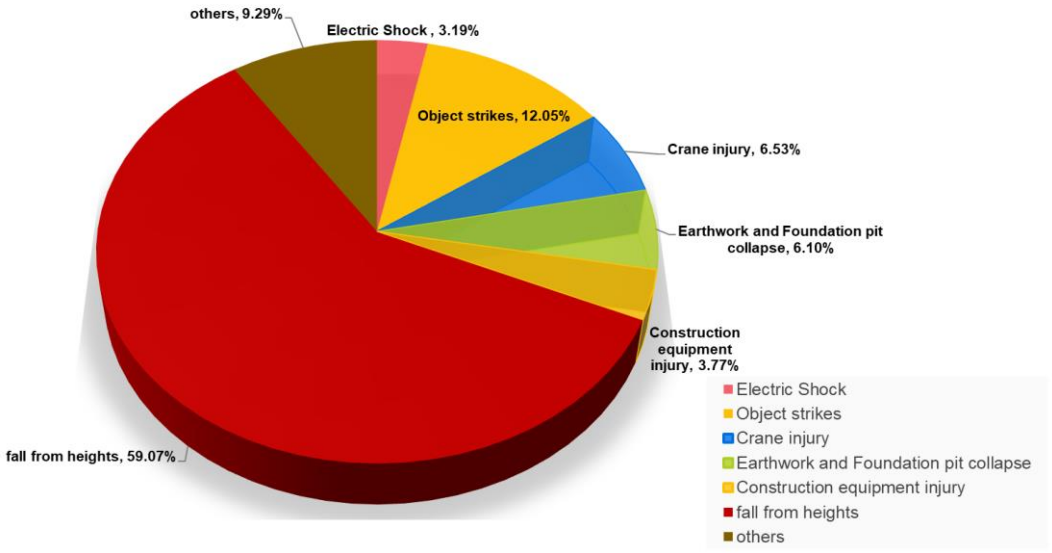


Figure 4 Safety Accident Types of Housing and Municipal Engineering Projects in China Mainland (2020)

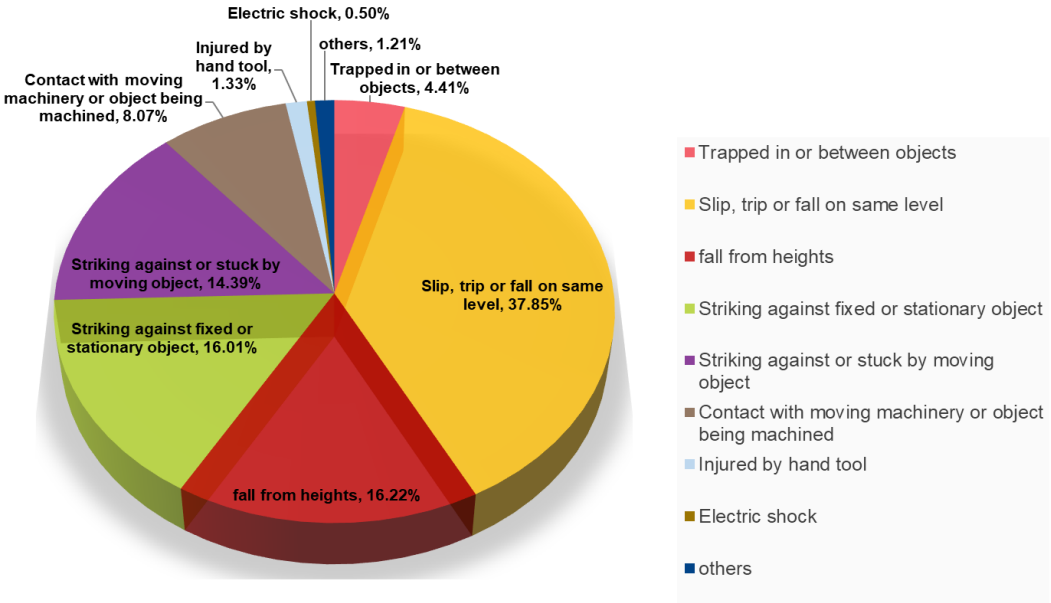


Figure 5 Safety Accident Types of Construction Industry in Hong Kong (2023)

Edge protection serves as one of the critical safeguards against falls, however, temporary safety structures are vulnerable to failure. These failures can stem from structural deficiencies (e.g., corroded railings with insufficient load capacity), managerial oversights (e.g., failure to reinstall protections after formwork removal), or environmental factors (e.g., strong winds displacing barriers) (Li *et al*., 2019). The dynamic environment within the construction process plays a significant role in fatal hazards. Typically, common protective equipment such as guardrail systems may quickly become noncompliant after installation (Johansen *et al*., 2024). The fast-paced nature of high-rise construction exacerbates these risks, as multiple trades working concurrently can lead to frequent breaches in safety barriers (Sulowski, 2014; Newas *et al*., 2022).

Compounding the issue, manual inspections are typically used to assess potential hazards and verify compliance with industry safety standards. Although safety codes (e.g., ISO 45001) for work at height specify minimum railing heights and load capacities, reliance on manual inspections introduces a multitude of challenges that can significantly undermine operational efficacy, especially given intermittent inspections and complex dynamic working conditions.

One of the primary challenges associated with manual inspection processes is the inherent variability in human performance. Factors such as fatigue, distractions, or inadequate training can significantly impact an inspector’s ability to identify defects or

anomalies consistently. This inconsistency not only increases safety hazards but can also result in costly rework or, in more severe cases, accidents. Moreover, the subjective nature of human assessments can lead to discrepancies in inspection outcomes, complicating efforts to maintain standardized safety benchmarks. As a result, hazards can persist unnoticed, highlighting the urgent need for automated, real-time monitoring systems that identify risks before accidents occur.

With this in mind, emerging technologies such as Building Information Modeling (BIM), Computer Vision (CV), and the Internet of Things (IoT) provide promising ways to automatically identify hazards and improve safety performance in construction. For instance, deep-learning-based computer-vision algorithms (i.e., the YOLO series models) have been widely used for safety hazard detection (e.g., identifying workers not wearing hard hats or safety harnesses) with detection accuracies exceeding 90% under controlled experimental conditions (Fang *et al*., 2018; 2020). Similarly, BIM-driven approaches have also been proposed to manage safety inspection information for better decision-making. BIM-based methods are generally involved in the early stages of the project and are widely used to identify hazards through comparisons between the as-built and as-planned models. Some scholars also used them to examine the defects of the building surfaces. For instance, Tan *et al*. (2022) focused on external walls and proposed a hybrid method for the automatic detection and localization of defects. By integrating CV, UAV, and BIM technology, the identified defects along

with detailed component information can be successfully mapped to the corresponding locations in BIM models.

Despite their success, there are two major limitations that hinder real-world applications on construction sites. First, existing deep-learning-based artificial intelligence monitoring systems have complex architectures and nonlinear transformations, often referred to as "*black boxes*," causing difficulty in understanding how AI systems arrive at their outputs. As a result, it is challenging for safety managers to trust the results generated by current AI systems in practical sites. Second, prior approaches rely mainly on 3D reconstruction or 2D image-based inspection for identifying hazards in construction, particularly in detecting missing or inadequate safety protections. Consequently, these methods cannot reliably determine the precise location of a detected hazard within the BIM/3D model or associate it with specific building components, limiting their usefulness for actionable decision-making.

Moreover, a close examination of these studies reveals that attention has focused mainly on the detection of hazards, with far less emphasis on predictive solutions based on past behavioral data of the workers (Kong *et al*., 2021). Traditionally, predictions of unsafe behavior have relied on root-cause models and psychological theories such as the Theory of Reasoned Action (TRA) and the Theory of Planned Behavior (TPB) (Sheppard *et al*., 1988; Guo *et al*., 2016), which explain *how* behavior changes under

different working conditions. These models assume that behavior is intentional; hence, they predict deliberate behavior (Ajzen, 1999).

However, such models are "widely applied without sufficient attention paid to what makes [them] work in [their] contexts of origin, and without adequate customization for the specifics," leading to a "flawed reductionist view" of safety issues (Peerally *et al*., 2017). In a similar vein, machine learning approaches such as artificial neural networks (ANN) have been used to predict unsafe behavior (Patel & Jha, 2015; Davoudi *et al*., 2019; Ayhan & Tokdemir, 2020). Yet, a substantial subset of these studies builds models from secondary evidence (e.g., meta-analyses within literature reviews) rather than operational site data; as a result, such models inherit heterogeneous definitions, reporting biases, and a lack of situational variables—factors that undermine their practical relevance on active construction sites.

Given these considerations, there is an urgent need to develop an explainable, AI-driven approach that reliably detects missing or inadequate safety protections and identifies workers approaching unprotected edges in real time from images/videos. Leveraging UAVs provides the site-level coverage and vantage points necessary for continuous monitoring, while explainability ensures that safety managers can audit model decisions and act with confidence-together enabling proactive hazard prevention on high-rise construction sites.

1.2 Research Aims and Objectives

To address these challenges, the study aims to investigate the following research question: *How can we improve the accuracy and interpretability of deep learning models in monitoring Fall-From-Height (FFH) risk in construction sites by using Unmanned Aerial Vehicles (UAVs) and computer vision systems*? To answer this research question, the study pursues the following specific research objectives:

- **Objective 1:** To examine current studies related to the workforce safety issue-specifically, falls from height risk in the construction sector, and the research progress of computer vision applications and explainable artificial intelligence (XAI) in construction, identifying the research questions, aims, and objectives that we need to address in this thesis.

- **Objective 2**: To develop a UAV-based approach that automatically and accurately identifies guardrails for safety protection, explicitly addressing multi-scale variations in UAV imagery to enable multi-scale feature perception with both local and global context, thereby improving detection accuracy.

- **Objective 3:** To design an XAI-enhanced visualization approach that improves the interpretability of detection results produced by the deep learning model and maps the detection information to dynamic 3D models, enabling reliable determination of the precise location of detected hazards within the 3D reconstruction model and their association with specific building components, thereby supporting safety

managers in making informed decisions and taking timely actions.

- **Objective 4**: To design a computer vision-based predictive model that anticipates, at an early stage, when workers are approaching unprotected edges from video data. The model will explicitly incorporate worker-worker interactions to predict their trajectories, improve robustness and accuracy in multi-worker tracking, and enable proactive hazard prevention.

- **Objective 5:** To conduct a comprehensive evaluation experiment on real-world projects, quantifying the approach's accuracy and reliability (e.g., precision/recall) and assessing its practical impact on inspection efficiency and hazard prevention.

1.3 Research Significance

While detecting hazards can enable prompt intervention in construction, traditional human inspection often falls short because it relies primarily on information gathered after accidents occur. To address this gap, this research utilizes computer vision and deep learning models, enabling direct observation of hazards in real-time and a re-examination of TPB's relevance to construction.

For example, studies show that safety incidents often materialize during rework (Love *et al*., 2018), in which unplanned work can lead to unplanned behavior rather than the planned behavior assumed by TPB. Observing hazards in real time, therefore, allows us to challenge prevailing theories and develop theories that reflect practice in dynamic

work conditions. This work integrates UAV technology, computer vision, and XAI algorithms into a unified framework to address key gaps in existing studies. With this in mind, the advantages of this research are threefold.

Firstly, by leveraging the wide-area and flexible imaging capabilities of UAVs and combining Spatial Adaptive Feature Modulation with a Feature Refinement (FRFN) module for layer-by-layer feature refinement in the deep learning model, this approach markedly improves robustness in detecting guardrails across varying viewpoints and scales. This not only reduces the burden of manual inspections but also produces structured outputs that facilitate subsequent localization and association with 3D reconstruction models and alert integration, supporting continuous safety monitoring.

Secondly, this approach enhances transparency in automated hazard detection by using deep learning model and computer vision. By incorporating XAI technique (e.g., visualizing feature maps), the XAI-based computer vision safety monitoring system can demonstrate why a detection or alert was generated by the deep learning model. As a result, construction safety managers receive alerts accompanied by visual evidence and highlighted factors that influenced the outcome, thereby supporting timely and well-founded decisions.

Thirdly, the proposed unsafe-behavior prediction model estimates workers' trajectories and flags approaches to unprotected edges in real time, enabling managers to take

immediate action. Site teams can issue alerts to workers approaching unguarded edges on high-rise buildings and intervene as necessary. Such feedback also enhances workers' safety awareness and facilitates the timely correction of unsafe actions in the construction site.

Beyond the technical advances, the proposed computer vision system offers practical benefits. The proposed approach improves worker safety through proactive hazard identification using computer vision and UAVs, reduces operational costs by replacing labor-intensive inspections with automated monitoring, and establishes a benchmark for explainable AI in safety applications on construction sites. By bridging the gap between laboratory performance and on-site practical applications, this research delivers a transparent and practical safety solution for workers, safety managers, regulators, and the construction industry. Moreover, our proposed approach enables stakeholders to trust the outcomes produced by the AI system.

1.4 Structure of the Thesis

To achieve the research aims and objectives, this thesis adopts a mixed-methods design integrating a systematic literature review, data acquisition, model development, and controlled experiment design. Figure 6 presents the end-to-end workflow and illustrates how Chapters 2-8 accomplish Objectives 1--5.

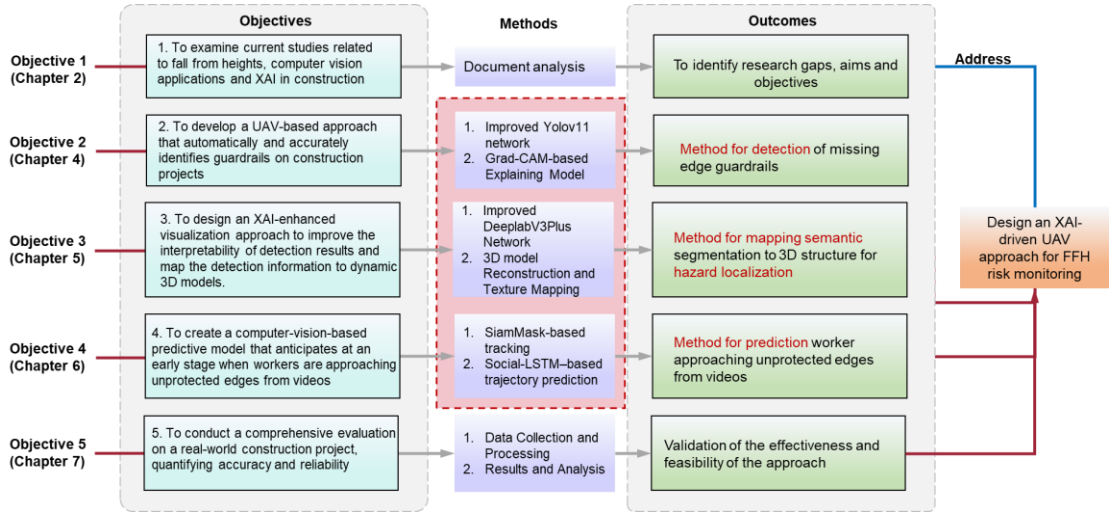


Figure 6 The overall framework of the study

1. Chapter 2 - *Literature review* **(Objective 1).** In this chapter, we examine current studies related to workforce safety issues—specifically, the prevention of falls from height (FFH) risks on construction sites—and review research progress in computer vision applications and explainable artificial intelligence (XAI) within the construction domain. Through this analysis, we identify research gaps, define research questions, and establish the aims and objectives of this thesis.

2. **Chapter 3 -** *Research design*. This chapter adopts the Design Science Research (DSR) paradigm to guide the conceptual framework design, model development, and evaluation. The DSR methodology consists of four key processes: (1) Problem identification and motivation; (2) Objectives of a solution; (3) Design and development; and (4) Demonstration.

3. **Chapter 4 -** *Improved interpretable computer vision for detecting guardrails for edge protection* **(Objective 2).** In this part, the improved YOLOv11 network architecture and the Grad-CAM-based explanation model are described in detail, along with model training and implementation procedures. Evaluation performance metrics, including precision, recall, and mAP, are also introduced and discussed.

4. **Chapter 5 -** *Mapping semantic segmentation to 3D reconstruction model for hazard localisation* **(Objective 3)**. Within this chapter, the improved DeepLabV3+ network architecture and its corresponding components, as well as the 3D model reconstruction and texture mapping processes, along with model training and implementation procedures, are discussed comparatively. Evaluation performance metrics, including precision, recall, mAP, and F1-score, are also introduced and comprehensively analyzed.

5. **Chapter 6 -** *Improved computer-vision learning model for risk prediction* **(Objective 4).** In this chapter, we describe the vision-based tracking and trajectory prediction models, along with the model training and implementation procedures. Evaluation performance metrics, including mAP, ADE, and FDE, are also presented and discussed in detail.

6. **Chapter 7 -** Experiments and Results **(Objective 5).** In this chapter, we present the case background, data collection and preprocessing procedures, along with results and analysis for (i) the detection and localization of missing edge guardrails, and (ii) the identification of workers approaching edges with missing guardrails.

7. **Chapter 8 -** Discussion and Conclusion. In this chapter, we summarize the main research findings and contributions of the thesis, including both theoretical insights and practical implementations. The limitations of the study and suggestions for future research directions are also discussed.

CHAPTER 2 LITERATURE REVIEW

2.1 Fall from Heights (FFH)

The construction industry has been blamed as one of the most hazardous among all industries, with daily dangers amplified by its complex and dynamic operations. Statistics indicate that approximately 50% of accidents are construction-related in both China and the United States (Uddin *et al*., 2020). These accidents are commonly categorized into five types: falls from heights (FFH), collapses, struck-by-object incidents, lifting injuries, and electrocutions (Kang, 2022). Among these, FFH is one of the most common causes of injuries and fatalities and has therefore attracted substantial attention among researchers and practitioners.

Against this backdrop, early identification of FFH hazards is critical for effective safety management. Working at height presents inherent risks, which is why many countries impose strict protections. In China, the *Technical Code for Safety of Working at Height of Building Construction* mandates protective measures for leading edges with a fall risk of 2 meters or more, vertical openings with short edges less than 500 mm, and non-vertical openings greater than 25 mm (MOHURD, 2016). Similarly, in the United States, OSHA defines a “hole” as a gap or void 5.1 cm or more in its least dimension in a walking/working surface. Consistent with these definitions, Na *et al*. (2021) identify falls from roofs, ladders, scaffolds, openings, and edges as the most frequent FFH events. These regulatory thresholds and empirical studies underscore the need for methods that can locate and characterize FFH hazards before accidents occur. These

insights motivate proactive solutions that intervene before risks materialize.

One of the prominent proactive solutions leverages building information modeling (BIM) to support hazard identification. For example, 4D BIM links 3D geometry with the construction schedule to visualize time-dependent risks. Integrations of BIM with augmented reality and location tracking enable real-time risk detection (Park & Kim, 2013), and BIM-based frameworks support design-phase hazard assessment (Rafindadi et al., 2022). Recent advances include automated FFH risk scoring (Kim *et al*., 2020), machine learning-aided path simulations (Yan *et al*., 2024), and plug-ins that connect BIM models to safety databases (Lu *et al*., 2021). Nevertheless, many BIM-centric approaches still demand substantial manual input and computation, limiting their scalability in dynamic working conditions.

On site, the absence of fall-protection measures remains a major contributor to FFH incidents. Manual hazard identification is error-prone under dynamic conditions and often relies on incomplete models. Rule-based BIM checking attempts to classify hazardous elements (e.g., openings, walls, rooftops) (Zhang *et al*., 2015; Pham *et al.*, 2020), yet such rule-based approaches often overlook safety-critical attributes within each category (Qi *et al*., 2014). Spatial approaches—such as compliant edge detection (Zhang *et al*., 2013) and voxelized point-cloud analysis (Wang *et al*., 2015)—offer greater precision but can be computationally intensive.

Recent studies have begun to enable real-time guardrail inspections by integrating computer vision and IoT technologies (Tariq *et al*., 2023; Johansen *et al*., 2024). These advancements expand FFH-prevention capabilities. For example, deep learning models can be used to detect missing guardrails (Kolar *et al*., 2018) and unsafe worker behaviors (Khan *et al*., 2021), 3D point clouds are used to facilitate verification of scaffold compliance (Wang *et al*., 2019), and ontology-based knowledge graphs (Fang *et al*., 2020) combined with semantic reasoning approaches (Wu *et al*., 2021) are employed to support automated hazard identification and mitigation.

Despite their success, most 2D image-based computer vision approaches are conducted in limited predefined scenarios and lack sufficient accuracy in detecting guardrails under multi-scale variations in UAV imagery, while 3D point-cloud-based methods typically involve heavy computational demands. Moreover, existing AI systems for guardrail detection often fail to reliably register detected objects within 3D reconstruction models or associate them with specific building components. Accordingly, there is an apparent necessity for an automated and computationally efficient method that (1) accurately detects guardrail conditions across varying scales in UAV imagery, and (2) precisely localizes each detected guardrail within the 3D model to support timely and informed on-site safety management.

2.2 Understanding Unsafe Behavior in Construction

Understanding the causes and mechanisms of unsafe behavior is central to safety

research in construction (Fogarty & Shaw, 2010; Patel & Jha, 2015; Love *et al*., 2018; Choi *et al*., 2020). The Theory of Reasoned Action (TRA) posits that target behaviors are under volitional control—i.e., people believe they can act whenever they choose (Spielberger, 2004). While TRA predicts volitional behaviors, its explanatory power is limited in contexts with constraints on construction sites. The Theory of Planned Behavior (TPB) (Ajzen, 1991) extends TRA by adding perceived behavioral control to account for internal and external constraints. Although TPB has deepened our understanding of unsafe behavior in construction, it tends to provide a retrospective account of actions and may be vulnerable to hindsight bias.

Complementing these psychological perspectives, researchers often analyze causal pathways using accident reports and risk-propagation models to explain why FFH incidents persist (Macrae, 2022; Qi *et al*., 2022). Traditional studies rely on structured data and statistical tests (e.g., chi-squared, regression), whereas unstructured text analysis can reveal human, equipment, environmental, and managerial factors with greater nuance (Zhou *et al*., 2023). To reduce manual and subjective coding, a variety of automated approaches have been employed. For example, a construction accident causality system using grey relational analysis (Zhang *et al*., 2020) and a model based on cognitive reliability and complex network theory (Liu *et al*., 2020) have been proposed.

Hybrid AI methods such as Bayesian networks are utilized to further capture

interdependencies and support quantitative risk assessment (Qi *et al*., 2022). Yet, despite these advances, empirically validated methods that predict unsafe behavior in real time remain limited. Accordingly, the next section introduces our approach, which aims to transition from retrospective explanation to prospective, site-ready inference.

2.3 UAV and Computer Vision for Safety Inspection in Construction

Due to the complex and dynamic nature of construction environments, advanced information technologies are increasingly being used for safety inspections, replacing traditional manual methods. Among these technologies, computer vision (CV) has gained prominence in construction health and safety because it enables non-intrusive, automated analysis that approximates human visual perception (Ma *et al.*, 2022). A variety of computer vision and deep learning algorithms have been proposed to automatically identify unsafe behaviors from surveillance video, thereby reducing reliance on continuous human monitoring.

Computer vision approaches are used to detect objects, determine spatial relationships among objects, track object movements, and identify worker and equipment activities on construction sites from videos and images (Liu *et al*., 2021). A key advantage of using computer vision for on-site safety management is its ability to automatically extract information about multiple objects in complex, dynamic environments without requiring physical tags or sensors (Fang *et al*., 2018; Chan *et al*., 2021). Together, these capabilities form the foundation of CV-based safety analytics, which identify unsafe

conditions by monitoring object categories, locations, trajectories, and interactions (Liu *et al*., 2022).

A considerable body of scholars operationalizes these capabilities to identify construction components, hazards, and safety non-compliance. Object detection and semantic segmentation algorithms (e.g., YOLO, Faster R-CNN, and Mask R-CNN) are commonly used to detect workers' unsafe behaviors (Fang *et al*., 2018, 2019; Fang *et al*., 2020). Among these, YOLO, a representative one-stage detector, performs end-to-end inference in a single forward pass, enabling real-time performance with relatively low computational requirements.

Variants of YOLO have been extensively adopted for enhancing safety at practical construction spots. For instance, Zhu *et al*. (2017) combined a YOLO-based detector with a Siamese network to identify and track workers, while Kim *et al*. (2019) applied YOLOv3 to detect workers, excavators, and wheel loaders in UAV-captured videos. By integrating an object detection approach with image rectification, Kim *et al*. (2019) further attempted to measure distances between workers and equipment and then issue alerts for potential struck-by incidents. In related work, Kim and Chi (2019) encoded sequential visual patterns and operation cycles using CNNs with a two-layer LSTM, successfully recognizing excavator actions and explaining operational patterns. Similarly, computer vision methods have been employed to monitor the positions of workers, equipment, and components to verify compliance with safety distance

requirements and ensure correct placement away from hazards (Chi *et al*., 2009; Neuhausen *et al*., 2018).

Beyond fixed cameras, unmanned aerial vehicles (UAVs) extend the reach of CV and are increasingly adopted for site planning, safety inspections, and productivity analysis. For example, drones equipped with high-resolution cameras and LiDAR can rapidly detect structural defects and identify hazards such as unstable scaffolding or unauthorized personnel in restricted areas (Martinez *et al*., 2020; Wang *et al*., 2019). Advanced UAV systems integrate AI algorithms to automatically identify safety violations, including missing guardrails or the absence of personal protective equipment (PPE).

Despite existing advances, a key limitation hinders the practical implementation in the detection of guardrails for edge protection, particularly under multi-scale variation in UAV imagery and when precise localization within a 3D reconstruction model is required. The next section outlines the details of our proposed approach for accurate detection and 3D reconstruction-consistent localization of guardrail-related FFH hazards.

2.4 Object Detection and Segmentation in Construction

Computer Vision (CV) and deep learning algorithms have gained extensive applications in construction domains for object detection due to their strong performance in

construction-related detection tasks. For example, Fang *et al*. (2018) used a Region-based Convolutional Neural Network (R-CNN) to detect people and excavators with accuracies of 91% and 95%, respectively. An *et al*. (2021) developed a multi-scale dataset with 13 categories (e.g., workers, tower cranes) and benchmarked state-of-the-art detectors (e.g., YOLOv3, Faster R-CNN), comparing methods in the construction contexts.

Detection architectures in construction are mainly classified into two types: (1) single-shot; and (2) two-stage. Here, single-shot models such as SSD (Liu *et al*., 2016) and YOLO (Redmon *et al*., 2016) offer high throughput suitable for edge devices, whereas two-stage models such as Faster R-CNN (Ren *et al*., 2015) typically achieve higher accuracy at greater computational cost. For example, Luo *et al*. (2020) leveraged YOLOv3 to identify people and plant status (moving vs. stationary) to determine stuck-by accident risk. Roberts and Golparvar-Fard (2019) applied RetinaNet to excavators. Luo *et al*. (2019) used YOLOv3 for worker detection, and Luo *et al*. (2018) employed Fast R-CNN to recognize 22 object classes. Existing studies highlight a consistent speed-accuracy trade-off when selecting detectors for real-time safety applications.

A series of deep learning-based detection approaches have been proposed in the computer science domain (Girshick *et al*., 2014; He *et al*., 2015; Lin *et al*., 2017; Wang *et al*., 2022; Guo *et al*., 2024). Within the YOLO line, YOLOv4 (Wu *et al*., 2022) significantly improved the accuracy in the detection of concrete cracks by utilizing

EvoNorm-S0 to minimize model parameters. Sun *et al*. (2024) improved anchor design and localization training loss in YOLOR-BDIM, achieving a 2.7% mAP improvement over the baseline. YOLOv7 introduced a feature enhancement module (FEM) to improve detection performance under multi-scale conditions (Ye *et al*., 2023), and YOLOv8 refined the detection head and parameter sharing mechanisms to better detect small objects while reducing computational complexity (Jocher *et al*., 2023). Nevertheless, the detection of small objects in large, high-resolution site imagery remains challenging, especially for thin or distant objects related to safety in construction. A brief overview of research applying object detection in construction can be seen in Table 1.

Compared to object detection, semantic and instance segmentation are used to detect hazardous areas at the pixel level (Chen *et al*., 2018). A plethora of similar studies have been undertaken and proposed. For example, U-Net employs an encoder-decoder with skip connections to preserve fine details (Ronneberger *et al*., 2015). Building on this line, Lu *et al*. (2024) proposed a hybrid U-Net with a Swin Transformer approach to detect infrastructure cracks. Yuan *et al*. (2022) proposed Shift Pooling PSPNet, replacing the original pyramid pooling module to better capture local features at grid edges. Fu *et al*. (2021) introduced distributed spatial pyramid pooling to enrich feature extraction. A brief overview of research applying object segmentation in construction can be seen in Table 1.

Despite their success, accurate segmentation of high-resolution images remains challenging due to memory limits and detail loss from aggressive downsampling. One effective solution is a patch-based tiling strategy using DeepLabv3+. In DeepLabv3+, images are divided into smaller, overlapping patches for local feature extraction, and then the features are mosaicked together. This approach mitigates downsampling artifacts and consistently improves segmentation accuracy at scale—particularly for thin or distant elements.

On construction sites, detectors used for safety monitoring should sustain real-time operation while remaining robust to multi-scale variations (e.g., UAV imagery) and class imbalance typical of site scenes. Although existing detectors have achieved satisfactory performance, detection is not able to provide pixel-level localization. In our subsequent sections, we integrate segmentation results and 3D registration to measure guardrail geometry and place detected hazards precisely within 3D models for risk visualization.

Table 1. Examples of Research on Object Detection and Segmentation in Construction

Task	**Methodology**	**Target Objects**	**Author**
Object Detection	You Only Look Once (Yolo)	People and excavator	Luo *et al*., (2020)
	RetinaNet	Excavator	Roberts & Golparvar-Fard (2019)

	Yolov3	People	Luo *et al*., (2019)
	Faster R-CNN	People and excavator	Fang *et al*., (2018)
	Fast R-CNN	22 classes of objects	Luo *et al*., (2018)
Object Segmentation	Fully Convolutional Network	Cracks	Yang *et al*., (2018)
	Mask R-CNN	Concrete surface bug hole	Wei *et al*., (2019)
	Mask R-CNN	Structural supports and people	Fang *et al*., (2019)
	Mask R-CNN	Floorboards, joists, air vents and pipework	Atkinson *et al*., (2020)

2.5 Vision-based Object Tracking and 3D Reconstruction in Construction

Visual object tracking estimates an object's trajectory across video by localizing its position frame-by-frame (e.g., workers and equipment). Classical approaches include point and kernel tracking, often coupled with particle filters (sequential Monte Carlo) to manage uncertainty (Kim *et al*., 2016; Zhu *et al*., 2016). These methods have been used effectively for multi-target tracking at construction spots (Bügler *et al*., 2017). However, when computer vision and deep learning are applied to people tracking in complex scenes, modern deep trackers consistently outperform point/kernel methods (Angah & Chen, 2020; Cai & Cai, 2020).

Most trackers (e.g., classical and deep) represent targets with axis-aligned bounding boxes, which can fail under cluttered backgrounds and strong scale variations (e.g., camera-target distance changes) (He *et al*., 2018; Wang *et al*., 2019). To address these limitations, mask-based tracking uses binary segmentation masks to capture object extents more precisely, improving robustness to occlusion and scale changes (Perazzi *et al*., 2017; Ci *et al*., 2018; Wang *et al*., 2019). The trade-off is a higher computational cost relative to box-based tracking.

To balance accuracy and speed, the fully convolutional Siamese family integrates real-time correlation matching with mask prediction. In particular, SiamMask augments SiamFC to deliver real-time tracking with segmentation, mitigating computational overhead while retaining fine object boundaries (Wang *et al*., 2019). Given SiamMask's state-of-the-art results on VOT-2018 and its throughput advantages, we adopt SiamMask as the backbone tracker in this research.

Accurately localizing unprotected edges on construction sites is challenging yet essential for guardrail analysis. Three-dimensional reconstruction and model visualization provide the spatial context needed to place hazards precisely. Deep learning-based techniques are increasingly supplanting traditional pipelines. For example, Chakraborty *et al*. (2017) proposed an SfM method that leverages CNN-based dense features. Laser scanning can deliver high modeling accuracy (Zhao *et al*., 2020; Bernardo *et al*., 2016), but costs and logistics constrain wide adoption. Classical 3D

methods can segment damage within defined regions of interest, yet accuracy often degrades outside those regions (Zeng *et al*., 2024).

Recent learning-based MVS models improve both efficiency and depth estimation accuracy, such as MVSNet (Yao *et al*., 2019; Gu *et al*., 2020), PatchMatchNet (Wang *et al*., 2021), and UCS-Net (Cheng *et al*., 2020). Notably, Wang & Gan (2024) integrate the PatchMatch concept into an end-to-end MVS framework with learnable, adaptive modules, achieving more than 2.5× speedups and 50% lower memory use relative to prior methods. However, models such as PatchMatchNet still require substantial training to generalize robustly across site conditions. For guardrail detection, we require an efficient, UAV-ready learning-based MVS pipeline that yields scale-robust reconstructions and supports precise 3D model registration of detected edges—enabling pixel-level hazard localization.

2.6 Explainable Artificial Intelligence (XAI) in Construction

Explainable artificial intelligence (XAI) refers to methods and systems that provide human-interpretable reasons for model decisions, thereby supporting understanding and trust (Miller, 2019). Interpretability spans intrinsically transparent models, such as rule-based systems, decision trees, generalized additive models (Lou *et al*., 2012), and post hoc tools for opaque models, including variable importance, partial dependence, and global sensitivity analysis (Friedman, 2001; Saltelli *et al*., 2006).

With abundant data and mature software frameworks, deep learning models have grown dramatically in size and reach (Lipton, 2018), increasing both their prevalence in safety critical settings (Geirhos *et al*., 2020) and their opacity (Doshi-Velez & Kim, 2017). High apparent accuracy can mask shortcut learning and spurious correlations, so correctness alone does not guarantee trustworthy predictions (Carvalho *et al*., 2019). In construction—where earlier rule-driven systems afforded traceability—modern perception-heavy workflows require modality-specific XAI to maintain accountability. This has accelerated robust post hoc techniques for class-discriminative localization, human-readable overlays, and pseudo-label seeding for weakly supervised segmentation (Forest *et al*., 2024; Love *et al*., 2023).

Images and 3D point clouds dominate site perception for inspection, progress management, and as-built documentation (Fang *et al*., 2020; Fang *et al*., 2020; Mirzaei *et al*., 2022). Images provide rich texture and semantics; point clouds provide accurate geometry and spatial context—together spanning the visual-geometric spectrum required for site understanding. XAI methods for these modalities can be grouped as follows. Based on how explanations are generated, XAI for these two data modalities can be grouped into the following categories:

- *Activation-/gradient-based attribution*: These methods propagate class-specific signals (activations, gradients, relevance rules) through the network to produce spatial attribution maps (Sundararajan *et al*., 2017). The CAM family (e.g., CAM,

Grad-CAM, Grad-CAM++, Score-CAM) remains widely used (Zhou *et al.*, 2016; Selvaraju *et al*., 2017; Chattopadhay *et al.*, 2018; Wang *et al*., 2020). Strengths include contiguous, semantically aligned regions; a key limitation is coarse localisation due to the lower spatial resolution of late feature maps (Ancona *et al.*, 2018). In construction, most applications focus on images, localising class-discriminative regions for materials/components (e.g., concrete, wooden structures, vacuum-insulated glazing) under image-level supervision, thus revealing the decision basis of CNN classifiers (Hacıefendioğlu *et al*., 2022; Riedel *et al*., 2022; Luo & Liu 2024). CAM variants have also been used in vision-language models for image captioning to identify ergonomic problems and solutions with visually interpretable evidence (Yong *et al*., 2024).

- *Perturbation-based (model-agnostic) methods*: These explain predictions by systematically modifying inputs—occlusion/masking, blurring, inpainting—and observing changes in outputs (Fong & Vedaldi, 2017). Representative tools include occlusion sensitivity, LIME (Ribeiro *et al*., 2016), RISE (Petsiuk *et al*., 2018), and SHAP (Lundberg & Lee, 2017). They make minimal assumptions about the model, enabling broad applicability, but are typically computationally intensive and can exhibit sampling variance; naïve masking may introduce out-of-distribution artefacts (Barredo Arrieta *et al*., 2020). In construction, applications primarily focus on time-series data (e.g., ground settlement, concrete performance) (Elhishi *et al*., 2023; Luo *et al*., 2024), with comparatively fewer studies on images/point

clouds. Examples include SHAP for grout volume prediction (integrating borehole images and construction parameters) and for quantifying geometric/spectral feature contributions in point cloud classifiers (Aksu & Demirel, 2024; Atik *et al*., 2024; Zhe *et al*., 2025).

- *Propagation and relevance methods*: These attribute predictions by back-propagating relevance rather than raw gradients or input perturbations. LRP redistributes the output score layer-wise under conservation rules, and DeepLIFT decomposes outputs relative to a baseline to yield per-pixel/per-point contributions (Bach *et al*., 2015; Shrikumar *et al*., 2017). Compared with gradient saliency or CAM, they can mitigate gradient saturation and produce sharper, less noisy maps (Ancona *et al*., 2018). However, they require white-box access (activations, gradients, layer ops), and outputs are sensitive to propagation rules/baselines (Montavon *et al*., 2018). Construction applications for images and point clouds remain limited relative to CAM and perturbation-based work.

- *Concept and example-based explanations:* These elevate explanations from pixels/points to human-aligned concepts or influential training examples. TCAV quantifies concept importance, while influence functions and TracIn identify impactful training samples (Koh & Liang, 2017; Kim *et al*., 2018; Pruthi *et al*., 2020). They provide semantically grounded or data-audit insights but require curated concept sets or access to training data/optimisation traces; approximations

can be model- and scale-dependent. In construction, techniques like t-SNE have been used to visualise hidden-layer embeddings and assess class separability (Geetha & Sim, 2022).

Next, we introduce the details of how to provide human readable rationales for detection and to inform 3D-consistent hazard localization.

CHAPTER 3 RESEARCH DESIGN

This thesis aims to explore the following research question: *How could we improve the accuracy and interpretability of deep learning models in monitoring Fall-From-Height (FFH) risk in construction sites by using Unmanned Aerial Vehicles (UAVs) and computer vision systems*? To this end, we adopted a *design science* research (DSR) paradigm to guide the conceptual framework design, model development, and evaluation. Following such a design science approach, we design, implement, and evaluate an XAI-enabled UAV-based computer vision system that (1) detects unprotected edges and guardrails, (2) precisely localizes hazards related to FFH within a three-dimensional (3D) construction site model and links the hazards to specific building components, and (3) predicts workers approaching unprotected edges from videos.

Design science seeks not only to describe, explain, and predict phenomena in the natural and social world, but also to design and evaluate solutions that improve human performance (Aken, 2004; Chu *et al*., 2018). In this sense, design science develops knowledge and applications that guide the design and implementation of approaches with organizational value (Geerts, 2011). Figure 7 presents the research process that we used to design and develop the proposed system. The implementation of design science includes the following four steps.

- *Problem identification & motivation*: Identifying the research problem and

recognizing the value of its solution to the problem.

- *Objectives of a solution*: Describing the problem to defining measurable, functional goals for the solution.
- *Design and development*: Establishing the functional requirements and building the model to describe the objectives.
- *Demonstration*: Demonstrating how well the developed model addresses the research problem and research objectives.

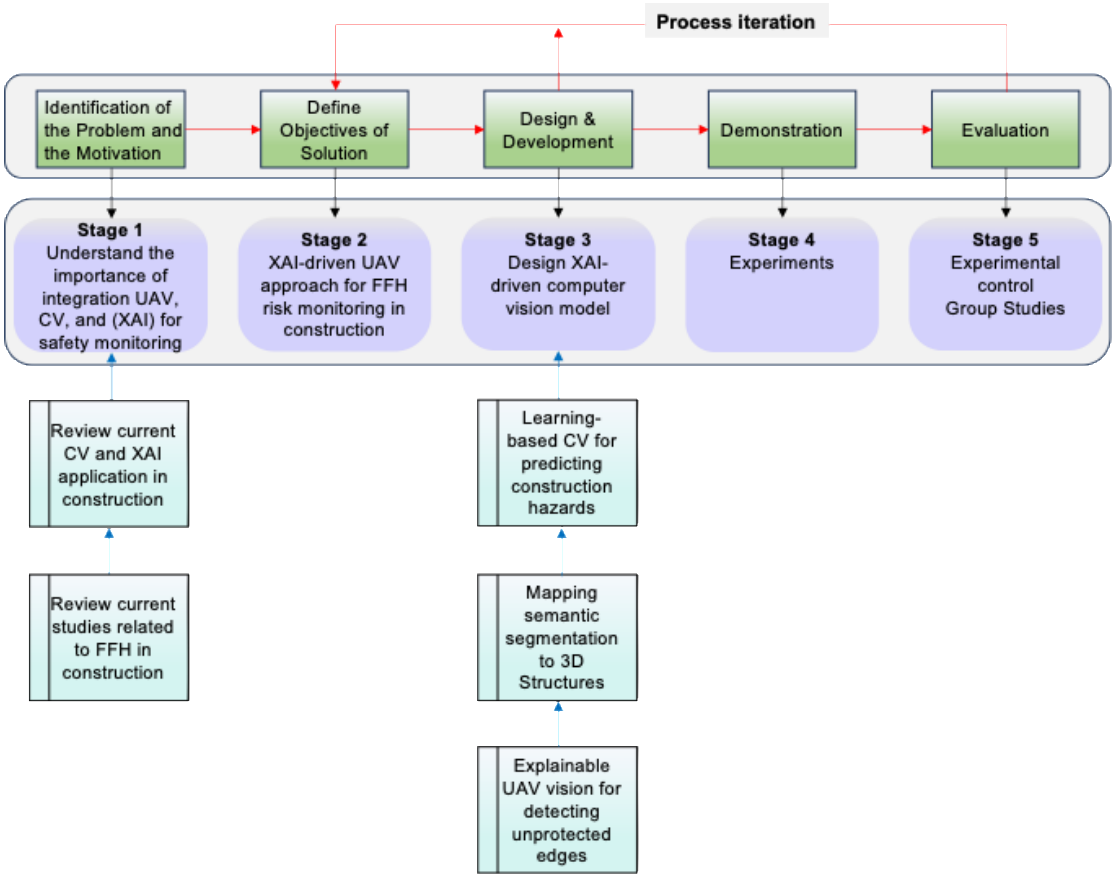


Figure 7 workflow of the design science approach

CHAPTER 4 AN IMPROVED INTERPRETABLE COMPUTER VISION APPROACH FOR DETECTING GUARDRAILS FOR EDGE PROTECTION

As discussed in Chapter 2, detecting guardrails is important for safety protection using computer vision and deep learning. Existing deep learning models are complex nonlinear statistical models that significantly suffer from the "black box" problem, making it difficult for on-site managers to understand why and how such AI systems produce their outcomes. Consequently, on-site managers are unable to trust the results generated by the AI system.

To address this limitation, we propose an improved interpretable computer-vision approach for detecting guardrails for edge protection in construction sites. Our proposed approach comprises two modules: (1) an improved YOLOv11-based detector for detecting guardrails and identifying their absence, and (2) a Grad-CAM-based explanation module used to highlight the visual evidence supporting each prediction. The basic framework of our proposed approach is exhibited in Figure 8.

In our improved YOLOv11 detector, spatial adaptive feature modulation (SAFM) makes the model capable of capturing multi-scale information, ranging from local to global levels, thereby reducing false positives and missed detections in complex backgrounds. The feature refinement feed-forward network (FRFN) refines feature maps layer by layer, improving discrimination of small objects under different working

conditions. In doing so, our improved YOLOv11 detector achieves high precision and recall in real-world construction projects. The details of each module of our approach are described in the following subsections.

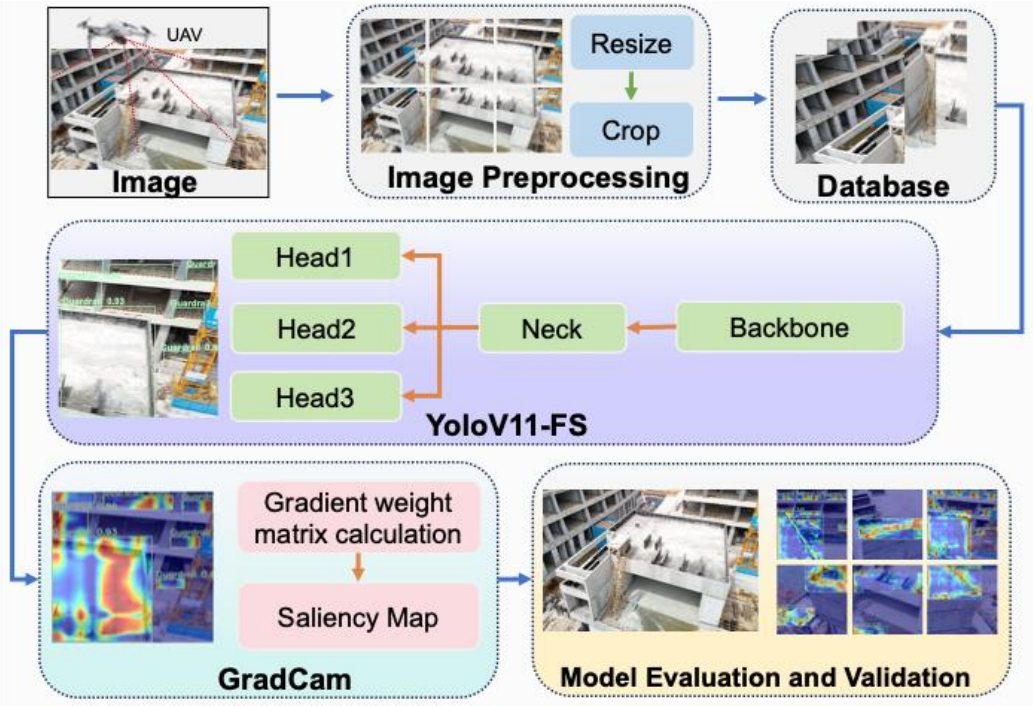


Figure 8 Workflow of our proposed approach

4.1 Improved Yolov11 network for Guardrail Detection

To accurately detect guardrails, we developed an improved detector named YOLOv11-FS. It builds upon the YOLOv11 architecture by (1) replacing the original C3K2 block with a Feature Refinement Module (FRFN), and (2) substituting the standard upsampling block with a Spatial Adaptive Feature Modulation (SAFM) block. These modifications enhance multi-scale feature representation and improve the detection of objects of varying sizes. The overall architecture of our proposed YOLOv11-FS is presented in Figure 9. Here, we first review the baseline YOLOv11 network and then describe the two optimization modules.

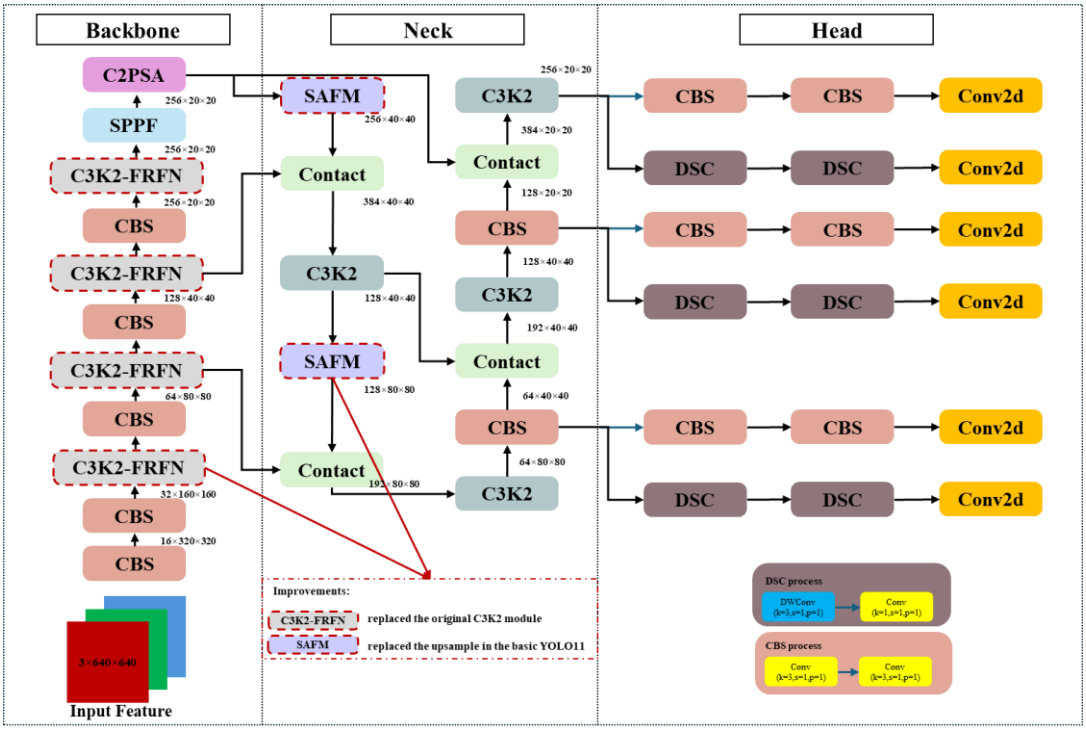


Figure 9 Network of Yolov11-FS

4.1.1 Network of Yolov11

YOLO (You Only Look Once) is a single-stage, deep learning-based object detector that formulates detection as a direct regression from images to bounding boxes and class probabilities. By eliminating region proposals and processing each image in a single pass, YOLO delivers real-time performance with end-to-end training in a unified framework. Owing to this speed-accuracy balance, the YOLO family is widely used in modern computer vision applications.

Building on this lineage, YOLOv11—released by Ultralytics in September 2024—advances both accuracy and speed. Its contributions focus on three aspects: (1) enhanced feature extraction via an optimized network architecture; (2) improved small-object detection through a dedicated augmentation module; and (3) dynamic sample

processing that adaptively identifies complex examples (e.g., occluded or tiny objects) and reweights the loss to prioritize their correct classification. By doing so, these changes enhance detection accuracy while maintaining real-time throughput.

Architecturally, YOLOv11 replaces the C2f blocks with more efficient C3K2 modules and incorporates SPPF for multi-scale pooling, which standardizes feature maps and enriches feature diversity. It also introduces Cross-Stage Partial with Pyramid Squeeze Attention (C2PSA), which extends the SE mechanism with a multi-level design to strengthen multi-scale feature processing. The model follows a three-stage design: (1) the Backbone extracts multi-scale features while gradually reducing spatial resolution; (2) the Neck upsamples, concatenates, and applies attention to fuse cross-scale information; and (3) the Head predicts bounding boxes, object confidence, and class probabilities while balancing speed and accuracy. The architecture of YOLOv11 is illustrated in Figure 10.

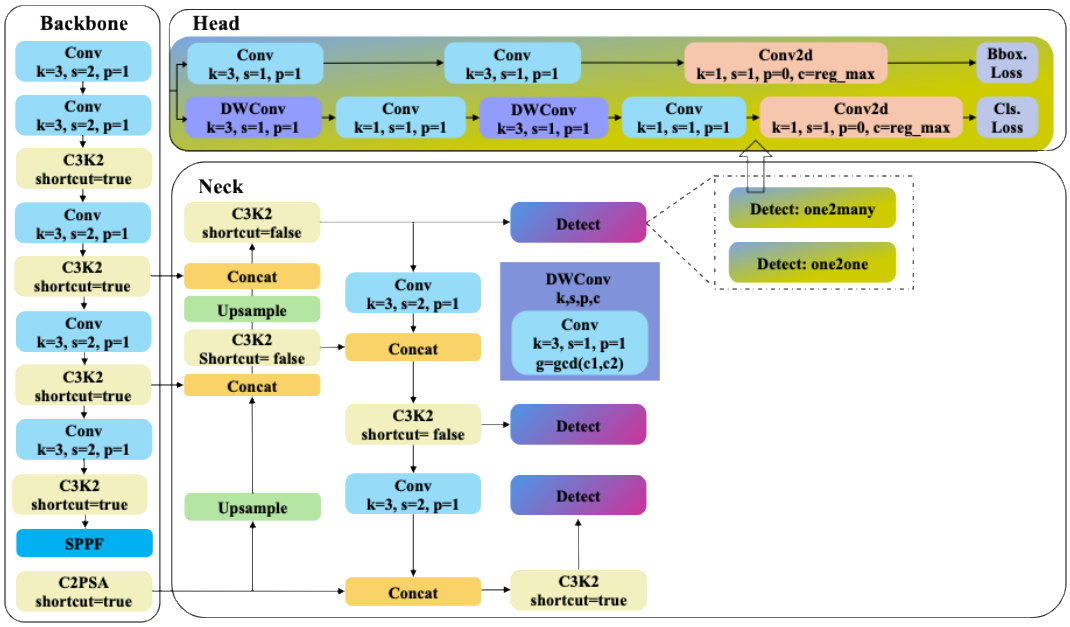


Figure 10 Network of YOLOv11

4.1.2 Spatial Adaptive Feature Modulation Module (SAFM)

Spatial Adaptive Feature Modulation (SAFM) is a spatially adaptive feature-modulation mechanism that dynamically modulates input features to select representative feature representations, thereby improving the restoration of image details and sharpness. The core of this mechanism is its ability to process features from a multi-scale perspective while incorporating both local and global information, resulting in more discriminative and robust features. The structure of SAFM is presented in Figure 11.

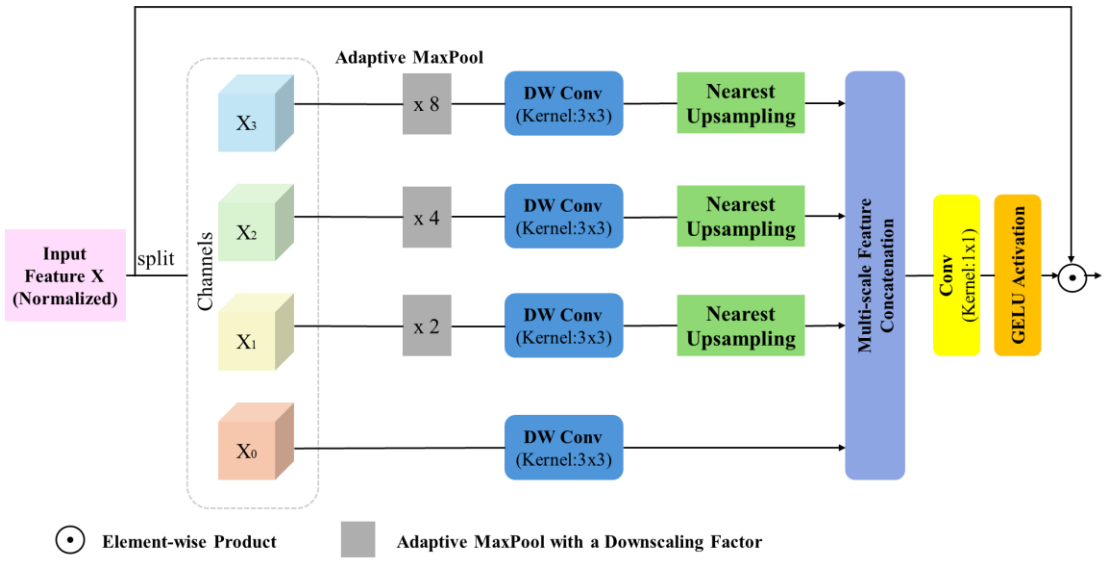


Figure 11 Structure of SAFM

The implementation of Spatial Adaptive Feature Modulation (SAFM) is primarily composed of the following three steps:

1) Channel Splitting and Preliminary Processing

First, the normalized input feature tensor is split along the channel dimension into four components. One of these components is processed using a 3×3 depthwise-separable convolution. The remaining three components are fed into a multi-scale feature-

generation unit, where operations such as downsampling, convolution, and pooling are applied to produce multi-scale features. The mathematical formulation of this step is as follows:

$$[X_0, X_1, X_2, X_3] = \text{Split(X)} \quad \text{Eq. [1]}$$

$$\hat{X}_0 = DWConv_{3\times3}(X_0) \quad \text{Eq. [2]}$$

$$\hat{X}_i = \uparrow_p (DWConv_{3\times3}(\downarrow_{\frac{p}{2^i}} (X_i))), 1 \leq i \leq 3 \quad \text{Eq. [3]}$$

Where, X represents the input features, $\text{Split}(\,)$ denotes the channel splitting operation, $DWConv_{3\times3}$ indicates a 3×3 depthwise separable convolution, $\uparrow_p$ signifies upsampling features of a specific level to the original resolution p through interpolation, and $\downarrow_{\frac{p}{2^i}}$ represents downsampling the input features to a size of $\frac{p}{2^i}$ via pooling.

2) Multi-scale Feature Integration

Next, the extracted multi-scale features are concatenated along the channel dimension and passed through a 3×3 convolution to aggregate both local and global relationships. This step can be formulated as follows:

$$\hat{X} = Conv_{1\times1}(Contat([\hat{X}_0, \hat{X}_1, \hat{X}_2, \hat{X}_3])) \quad \text{Eq. [4]}$$

where *Contat* denotes the concatenation operation, and $Conv_{1\times1}(\,)$ represents a 1×1 convolution.

3) Attention Estimation and Adaptive Modulation

After obtaining the aggregated features, a normalization layer followed by the GELU nonlinearity is applied to estimate the attention map. Finally, the original input features are adaptively modulated using this attention. This step is formulated as:

$$\bar{X} = \emptyset(\hat{X}) \odot X \qquad \text{Eq. [5]}$$

where $\phi(\cdot)$ denotes the GELU function, and $\odot$ represents element-wise (Hadamard) multiplication.

4.1.3 Feature Refinement Module (FRFN)

Feature Refinement Feed-Forward Network (FRFN) is designed to enhance feature-representation capabilities in image-processing tasks. Its core design principle is to progressively refine and optimize feature maps to achieve higher accuracy in classification and detection. The structure of FRFN is exhibited in Figure 12. The key implementation steps are described in detail:

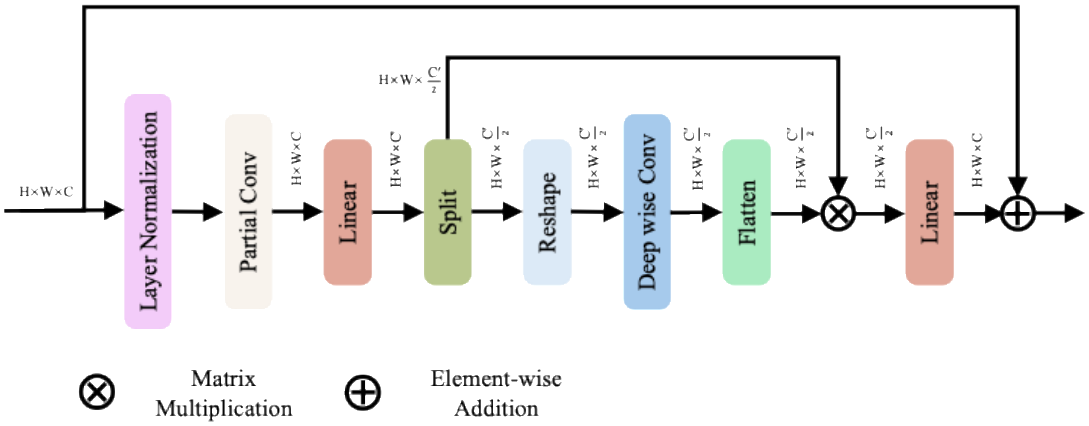


Figure 12 Structure of FRFN

The input features of FRFN first pass through a GELU activation and a linear projection ($W1$), followed by partial convolution to reinforce informative regions. The features are then split into two branches: (1) one branch undergoes depthwise convolution (DWConv), reshaping, and flattening to introduce locality; (2) the other branch interacts with the first via a gating mechanism to reduce redundant information.

Next, the processed features are projected through another linear layer ($W2$) and a GELU activation to produce the output. The mathematical formulation is as follows:

$$\hat{X}' = GELU\left(W_1 PConv(\hat{X})\right) \quad \text{Eq. [6]}$$

$$[\hat{X}'_1, \hat{X}'_2] = \text{Split}(\hat{X}') \quad \text{Eq. [7]}$$

$$\hat{X}'_{\text{r}} = \hat{X}'_1 \otimes F(DWConv(R(\hat{X}'_2))) \quad \text{Eq. [8]}$$

$$\hat{X}'_{\text{out}} = \text{GELU}(W_2 \hat{X}'_{\text{r}}) \quad \text{Eq. [9]}$$

Where, W_1 and W_2 are linear projection operations, Split() is the channel splitting operation, $F()$ and $R()$ are the reshape and flatten operations respectively, PConv () is partial convolution operation, DWConv() is depthwise separable convolution, and $\otimes$ symbolizes the cross product of matrices.

4.2 Grad-CAM-based Explaining Model

Grad-CAM visualizes the class-discriminative regions that a convolutional neural network (CNN) relies on when making a prediction. It does so by using the gradients of a target class flowing into the last convolutional layer, where the feature maps are most semantically aligned with the output.

Specifically, we compute the gradient of the target score with regard to the selected layer's feature maps and global-average-pool these gradients to obtain one importance weight per channel. The feature maps are then calculated by a weighted sum, followed by a ReLU, to produce a coarse class activation map that retains only features contributing positively to the target class. Subsequently, an upsample process and an overlap operation are conducted on this map, yielding an intuitive heatmap of the regions that drive the decision. Owing to its simplicity and interpretability, Grad-CAM provides a transparent link between model outputs and image evidence. The workflow of the Grad-CAM is shown in Figure 13.

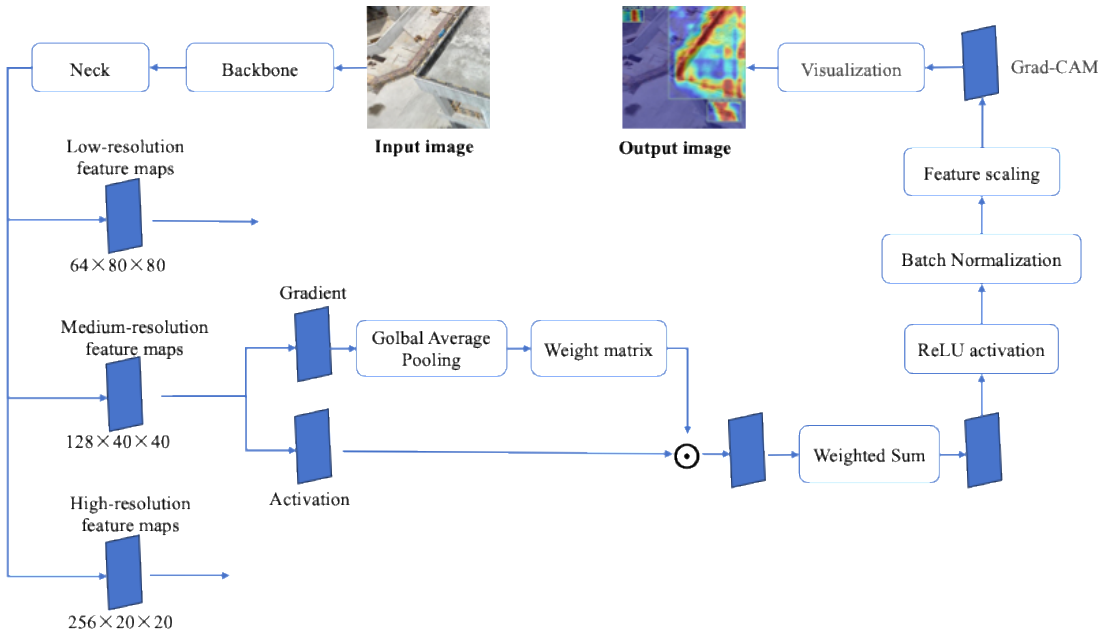


Figure 13 Workflow of Grad-CAM

The procedures for implementing Grad-CAM algorithm are as follows:

CAM replaces the fully connected layer after the last convolutional layer of a convolutional neural network with a Global Average Pooling (GAP) layer, which is connected directly to the output layer. In this way, the multi-channel feature maps produced by the convolutional layer are each pooled by the GAP layer into a single value in a one-dimensional feature vector, and the connection weight between this value and the output becomes the weight of the corresponding feature map. Using these weights to take a weighted sum over the feature maps across channels, then upsampling the result to the original image size and normalizing it, yields a weight matrix for the input image.

The computational formulas are as follows:

$$M_c(i,j) = \sum_{k=1}^{K} w_k^c \cdot A_k(i,j) \quad \text{Eq. [10]}$$

$$M_c^*(i,j) = \frac{M_c(i,j) - M_{c,min}}{M_{c,max} - M_{c,min}} \quad \text{Eq. [11]}$$

Where $M_c(i,j)$ is the element in row i column j of the (pre-normalization) weight matrix M_c when the sample is recognized as class ccc; k donated the total channels in the feature maps output by the last convolutional layer; w_k^c is the weight for class ccc associated with the k-th feature map; $A_k(i,j)$ is the element at row iii, column j of the k-th feature map in the last convolutional layer; and $M_{c,max}$ and $M_{c,min}$ are the maximum and minimum values of the (pre-normalization) weight matrix, respectively.

CAM demonstrated excellent performance in simple computations and parameter efficiency. By employing a Global Average Pooling (GAP) layer instead of a fully connected layer, the overfitting problem is reduced to some extent. However, CAM has an inherent limitation: it requires the network to employ a GAP layer to connect to the output, which necessitates retraining and greatly limits the method's applicability. Building on this, Grad-CAM improves upon CAM by using the gradients of the feature maps in place of the GAP outputs, allowing the method to be applied to neural networks of arbitrary architecture.

Grad-CAM integrates gradient information with class activation mapping. Here, it uses the gradients of the elements in each channel of the last convolutional layer to compute a weight for the corresponding feature map, and then forms a weighted sum across channels to obtain the image regions most closely related to the predicted target.

Specifically, for class c, Grad-CAM computes the weight w_k^c of the k-th feature map by averaging, over spatial locations, the gradients of the class score y^c with respect to the pixels of that feature map. The corresponding formulas are as follows:

$$w_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad \text{Eq. [12]}$$

Where y^c is the score for the target class c obtained from the forward pass; A_{ij}^k is the value at spatial location (i, j) in the k-th channel of the feature maps A; and Z is the product of the spatial dimensions (width × times× height), i.e., the total number of spatial elements.

Since the explainability approach aims to retain feature components that positively influence recognition of the target class while discarding features that are detrimental or irrelevant, the ReLU function is introduced to remove negative values in the weight matrix, preserving only features useful for that class and thereby improving the accuracy of localizing important regions. The corresponding formula is as follows:

$$M_{c,GradCAM} = Relu(\sum_k w_k^c A^k) \quad \text{Eq. [13]}$$

$$M_{c,GradCAM}^* = \frac{M_{c,GradCAM} - M_{c,min}^{GradCAM}}{M_{c,max}^{GradCAM} - M_{c,min}^{GradCAM}} \quad \text{Eq. [14]}$$

Where $M_{c,GradCAM}$ denotes the (pre-normalization) weight matrix; $M_{c,max}^{GradCAM}$ and $M_{c,min}^{GradCAM}$ are the maximum and minimum values of that pre-normalization matrix,

respectively; and $M^{*}_{c,GradCAM}$ is the normalized weight matrix.

4.3 Model Training and Implementation Details

This section presents the detail of the detector configuration, model training, and the implementation of the YOLOv11-FS module and Grad-CAM explanation module used to detect guardrails. Here, we first introduce the training loss, followed by optimization settings and hyperparameter selection.

The experimental platform runs Ubuntu 20.04 as the operating system, equipped with an Intel Xeon Gold 6138 processor, 128 GB of system memory, an NVIDIA RTX A5000 GPU, PyTorch 2.0.1, CUDA 11.8, and Python 3.8.0. The batch size is set to 16, the input image size to 448×448, the number of epochs to 100, patience to 20, the initial learning rate to 0.01, and the weight decay coefficient to 0.005. All experiments were trained and validated using the same set of hyperparameters.

The YOLOv11-FS loss function integrates three components: (1) the classification loss L_{cls}; (2) the bounding-box regression loss L_{box}; and (3) the distribution focal loss L_{dfl}. By weighting these terms and employing advanced optimization algorithms, the model's performance on object detection is effectively improved. The YOLOv11 loss is expressed by Eq. [15]:

$$L_{YOLOv11} = L_{cls} + L_{box} + L_{dfl} \quad \text{Eq. [15]}$$

Here, L_{cls}, based on cross-entropy, measures the discrepancy between the predicted class probability distribution and the ground-truth labels, enhancing multi-class recognition capability. L_{box} optimizes the geometric alignment between predicted and ground-truth boxes to ensure precise localization. L_{dfl} adopts a dynamic weighting strategy that emphasizes the loss contribution of hard examples, significantly improving the detection of sparse targets.

4.4 Evaluation Performance Metrics

This subsection defines the metrics used to evaluate the detection effectiveness and feasibility (i.e., accuracy). In our research, the confidence interval is set to 95%. Four key performance indicators (KPIs) are applied to evaluate the results performance: (1) *precision*; (2) recall; (3) *m*AP. Next, we introduce the details of each KPI.

(1) Precision

Precision (also called positive predictive value) is a metric used to measure *how* often a machine learning model correctly predicts the positive class. It can be calculated by dividing the number of correct positive predictions (true positives) by the total number of instances the model predicted as positive (both true positives and false positives). The *precision* (P) can be defined by Eq. [16]:

$$P = TP / (TP + FP) \quad \text{Eq. [16]}$$

where *TP* denotes true positives (the number of correctly detected targets) and *FP* denotes false positives (the number of incorrectly detected targets).

(2) Recall

Recall (also called sensitivity) is a metric used to measure *how* often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset. It can be calculated by dividing the number of true positives by the total number of actual positive instances, which includes both true positives and false negatives. The *Recall* (R) can be defined by Eq. [17]:

$$R = TP / (TP + FN) \quad \text{Eq. [17]}$$

where *FN* denotes false negatives, i.e., the number of targets that actually exist but were not detected.

(3) mAP

Mean Average Precision (mAP) is one of the commonly used KPIs to assess the performance of computer vision models in detection tasks. It is equal to the average of the Average Precision metric across all classes in a computer vision model. mAP can be used to compare different models on the same task as well as different versions of the same model, and its value ranges between 0 and 1. The *m*AP can be defined by Eq. [18] and Eq. [19]:

$$AP = \int_0^1 P(y)dy \quad \text{Eq. [18]}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \quad \text{Eq. [19]}$$

where AP_i denotes the average precision for the class indexed by i, and n represents the total number of classes for the dataset.

4.5 Summary

This chapter introduced an interpretable computer-vision approach for detecting guardrails for edge protection in construction projects. Traditional deep learning-based computer vision models can achieve high accuracy in the detection of guardrails, while providing little transparency about why and how a prediction was made, which hinders their application, as on-site managers are unable to trust the results generated by the AI system. To address this limitation, we propose an improved interpretable computer-vision approach for detecting guardrails for edge protection in construction projects. First, an improved YOLOv11-based detector is proposed to identify guardrails. Second, a Grad-CAM-based explanation module is proposed to highlight the visual evidence supporting each prediction.

In sum, this chapter detailed the motivation for interpretability in construction-site safety, described the architecture and implementation of the detector and explanation module, and outlined how the generated attribution maps can support model refinement. This interpretable design aims to increase managers' trust in the utilization of deep learning models, reduce risk errors, and facilitate effective collaboration between computer-vision systems to improve safety performance.

CHAPTER 5 MAPPING SEMANTIC SEGMENTATION TO 3D RECONSTRUCTION MODEL FOR HAZARDS LOCALIZATION

Chapter 5 demonstrates the accurate detection of guardrails, but it often struggles to precisely localize hazards within a 3D reconstruction model or associate them with specific building components. To address this limitation, we first apply an improved DeepLabV3+ network to semantically segment safety guardrails in site imagery. Next, the segmented masks are mapped onto a 3D reconstruction of the workspace via texture projection. By anchoring 2D predictions in 3D, the method enables precise localization of unguarded edges in the as-built 3D model. This section describes the details of the segmentation model, the image-to-model mapping procedure, and how the fused 3D semantics support hazard visualization and decision-making. The workflow of our proposed approach is presented in Figure 14.

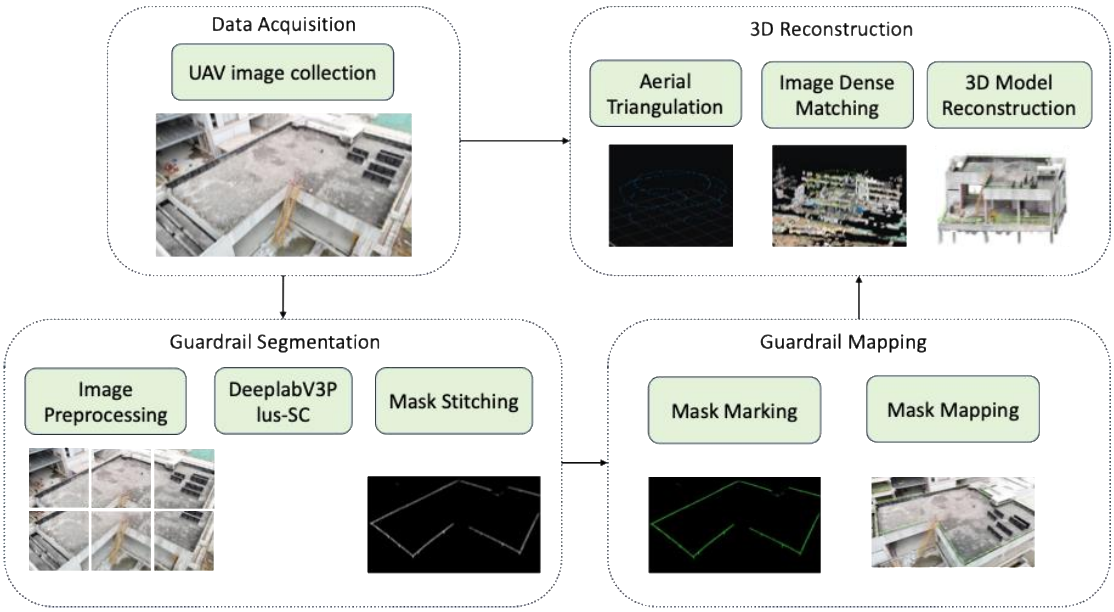


Figure 14 Workflow of our proposed approach

5.1 Improved DeeplabV3Plus Network

For images where guardrail areas are detected, an improved guardrail segmentation algorithm, DeeplabV3Plus-SC, is developed to accurately segment the guardrail objects. This proposed approach is based on the DeeplabV3Plus network architecture, with the addition of the convolutional block attention module (CBAM) in the encoder and the Squeeze-and-Excitation (SE) channel-attention module in the decoder. These enhancements improve the network's ability to process input features. The network structure is shown in Figure 15. Below are the details of the DeeplabV3Plus network and its optimization modules.

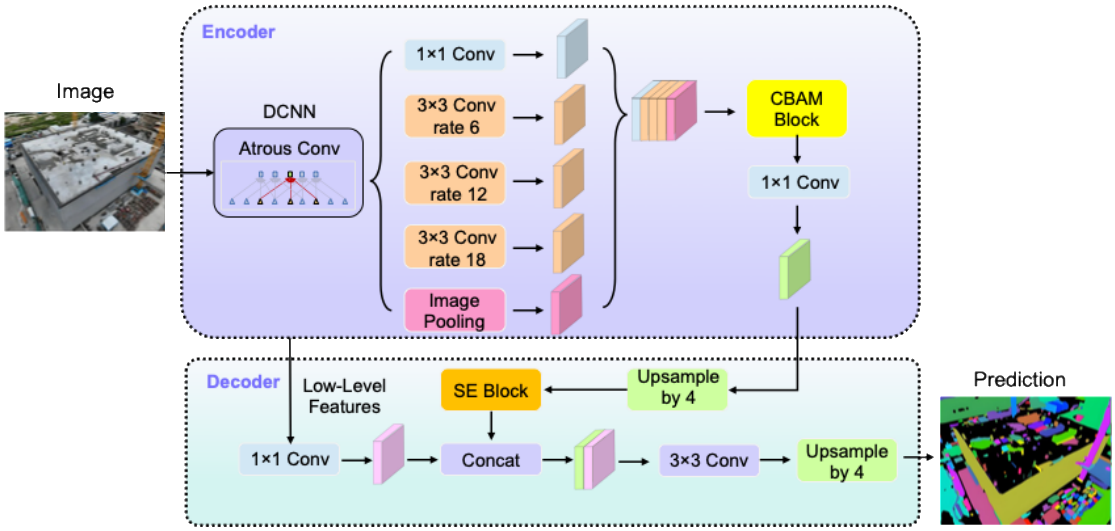


Figure 15 Network of DeeplabV3Plus-SC

Next, we present the details of the backbone DeeplabV3Plus semantic segmentation network and its optimization modules, including the Squeeze-and-Excitation (SE) channel-attention module and the Convolutional Block Attention Module (CBAM).

5.1.1 DeeplabV3Plus Semantic Segmentation Network

DeeplabV3Plus is an advanced deep-learning model for semantic segmentation tasks. The DeeplabV3Plus model is built on an encoder-decoder structure, in which the former

focuses on feature extraction from the images, and the latter enables these features to be mapped back to the original image size, achieving pixel-level classification. Specifically, the model's backbone network is responsible for feature extraction and includes two parts: (1) high-level semantic extraction and (2) low-level semantic extraction. The key component of DeeplabV3Plus is the ASPP module, which uses dilated convolutions to capture information at different scales. Features from different scales are combined and then merged with shallow features before upsampling to obtain the final prediction. The network structure of DeeplabV3Plus is demonstrated in Figure 16.

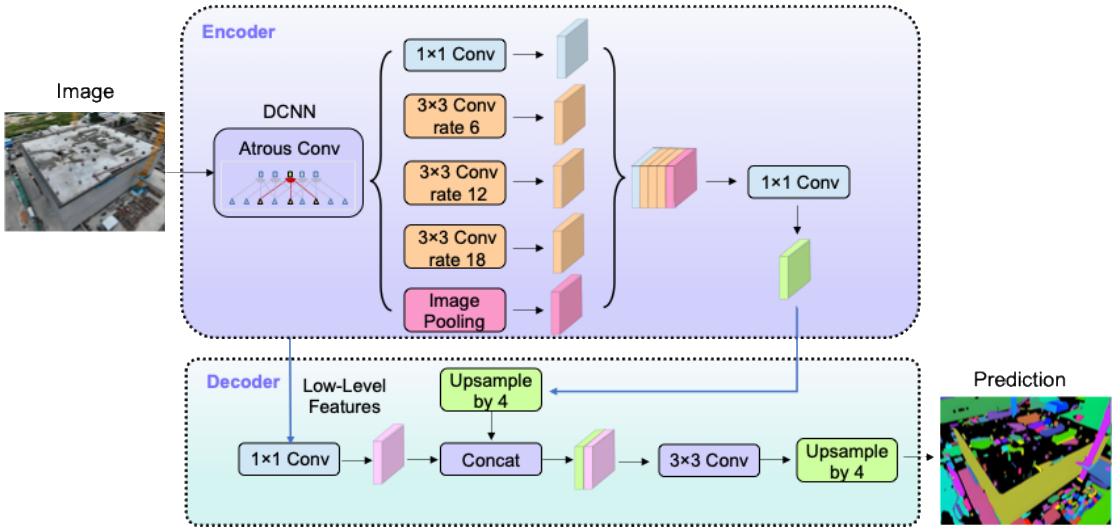


Figure 16 Network of DeeplabV3Plus

The procedure of the implementing DeeplabV3Plus is as follows:

Step 1. Backbone + ASPP

The original image passes through a backbone network (e.g., ResNet or Xception) for feature extraction, producing two outputs: (1) a shallow feature map (x1) and (2) a deep feature map (x2). The deep feature map (x2) is fed into the ASPP module, which has

five branches:

(1) Branch 1: 1×1 convolution (keeps spatial size).

(2) Branch 2: 3×3 convolution with dilation rate 6 (keeps spatial size).

(3) Branch 3: 3×3 convolution with dilation rate 12 (keeps spatial size).

(4) Branch 4: 3×3 convolution with dilation rate 18 (keeps spatial size).

(5) Branch 5: average pooling followed by a 1×1 convolution (adjusts channel number).

The outputs from all five branches are concatenated along the channel dimension.

Step 2. Feature fusion and classification

A 1×1 convolution is employed to obtain the final deep feature map (x3). The shallow feature map (x1) is processed with a 1×1 convolution to produce (x4). The deep feature map (x3) is upsampled to match the spatial size of (x1), yielding (x5). The shallow feature (x4) and the upsampled deep feature (x5) are concatenated along the channel dimension, and a 3×3 convolution is applied to produce the classification logits.

Step 3. Upsampling to image resolution

The logits are upsampled to the original input image size to produce the final segmentation result.

5.1.2 Channel Attention Mechanism SE Module

The *Squeeze-and-Excitation* (SE) channel-attention mechanism improves network performance with minimal additional computation by explicitly modeling dependencies between convolutional feature channels. The SE module consists of three operations:

(1) Squeeze, (2) Excitation, and (3) Scale. The structure of SE Module is shown in Figure 17.

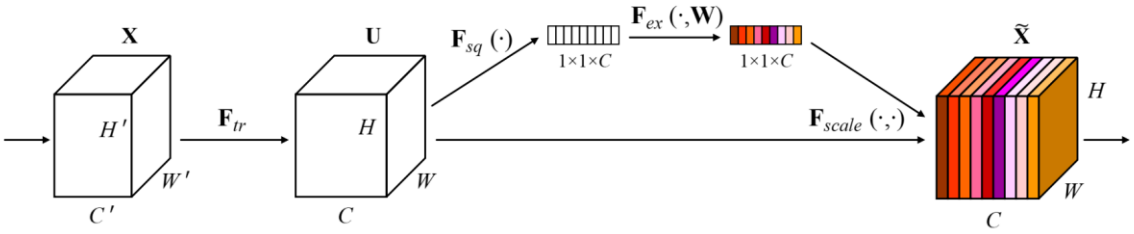


Figure 17 Structure of SE Module

Next, we introduce the details of three operations of the SE module.

(1) *Squeeze*. The SE module first aggregates the spatial dimensions of the input feature map using global average pooling, generating a channel descriptor for each channel. Within this step, global spatial information can be well compressed into a channel vector while the overall distribution of channel-wise feature responses is captured, which is of vital importance for the subsequent recalibration.

(2) *Excitation*. After squeezing, an excitation mechanism, essentially a self-gating module composed of two fully connected (FC) layers and a nonlinear activation, is adopted here. Begin with the FC layer, where the dimensionality of the channel descriptor can be effectively reduced. A ReLU activation is then applied, followed by the second FC layer that projects it back to the original channel dimension. A sigmoid activation then produces channel-wise weights. This process models nonlinear interactions between channels and yields a set of attention weights.

(3) *Scale*. The excitation output is used to recalibrate the original input feature map. Here, each channel is scaled by its corresponding weight, selectively emphasizing

informative features while suppressing less useful ones. This enables the model to focus on the most relevant features for the task.

5.1.3 Convolutional Block Attention Module (CBAM) Module

Convolutional Block Attention Module (CABM) is a lightweight attention module that enhances a CNN's ability to interpret input features. It focuses on relationships between channels and spatial locations, helping the network emphasize important information and suppress less useful details. By applying channel attention followed by spatial attention, the model automatically adjusts feature importance and improves feature representation quality. The structure of Convolutional Block Attention Module is displayed in Figure 18.

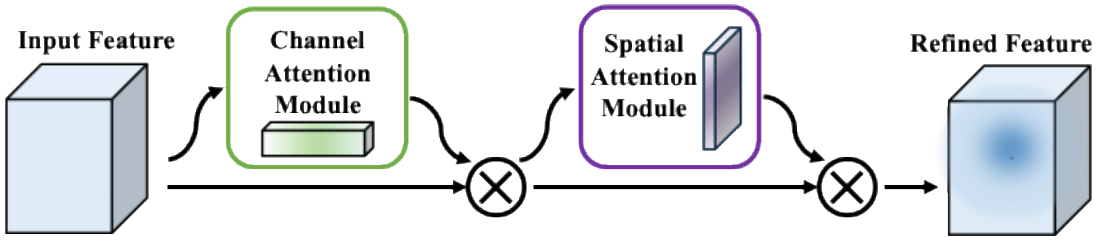


Figure 18 Structure of CBAM Module

The procedure of CBAM is as follows.

(1) The feature map FFF is fed into the module.

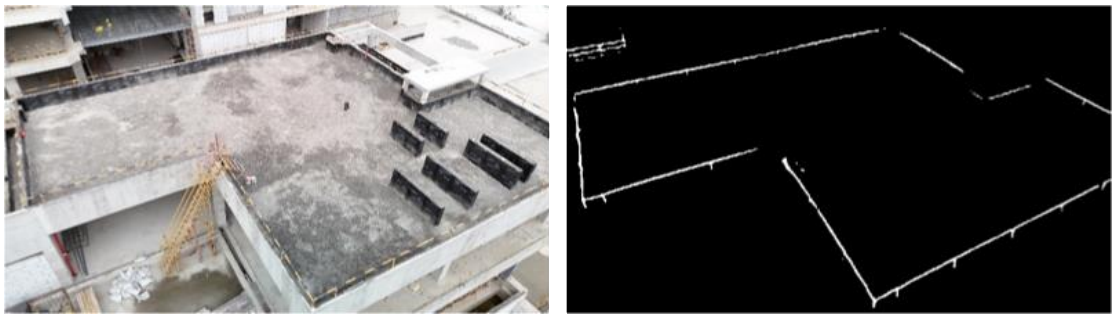
(2) The channel-attention submodule computes an attention map using a 1D convolution (MCM_CMC) and multiplies it with the current feature map to produce the channel-refined output.

(3) The resulting features are passed to the spatial-attention submodule, which computes a spatial attention map using a 2D convolution (MSM_SMS).

(4) The final output is obtained by multiplying this spatial attention map with the channel-refined feature map.

5.2 3D model Reconstruction and Texture Mapping

Using the improved guardrail segmentation algorithm, the collected images were segmented to extract edge guardrails, producing a binary segmentation result, as shown in Figure 19.



(a) Original image (b) Binary guardrail segmentation

Figure 19 Example of binary segmentation

Next, the binary segmentation result was re-mapped onto the original UAV images by assigning a specified color (green) and overlaying it, while preserving the original camera parameters and geolocation metadata, yielding a mask-mapped guardrail image.

Figure 20 Mask-mapped guardrail image

The mask-mapped images and the original images were then reconstructed in ContextCapture. The detailed steps are as follows:

Step 1: ***Aerial triangulation***. Based on the guardrail detection and segmentation results, both the original images (without guardrail overlays) and the mask-mapped images were used jointly as input imagery. Using the images' camera parameters and geolocation information, the software performed aerial triangulation to compute the relative poses between images, as illustrated below.

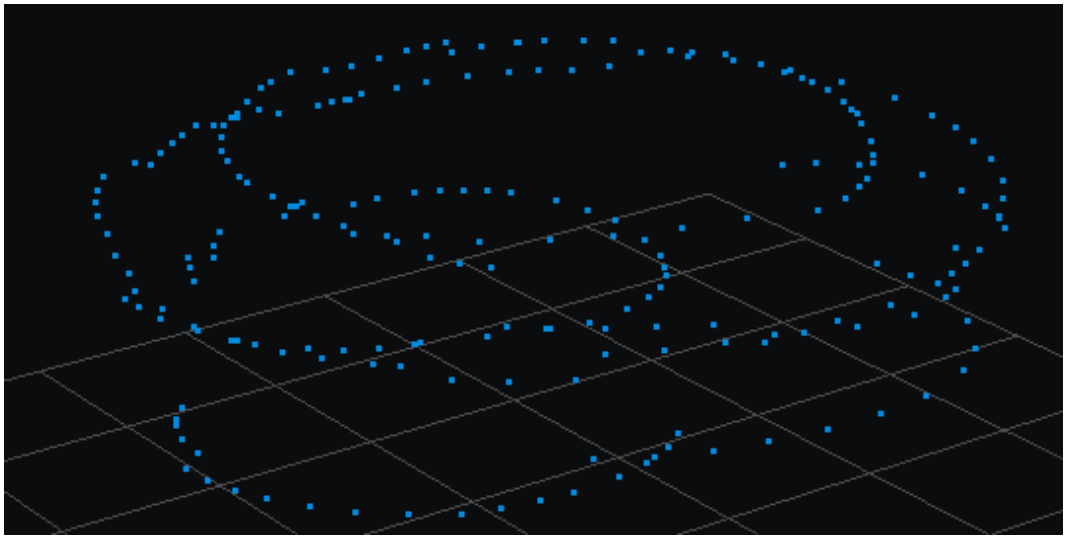


Figure 21 Relative poses between images

Step 2: ***Dense multi-view matching.*** Using the triangulation results, the software performed dense matching across multiple views. The matching output is shown below.

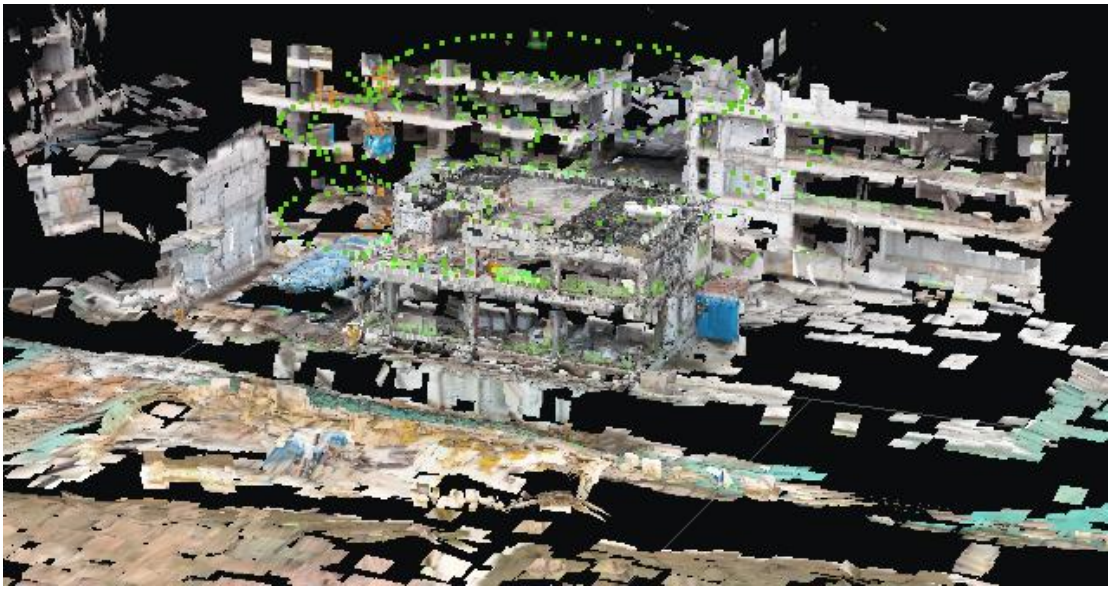


Figure 22 Results of *dense multi-view* matching

Step 3: ***3D reconstruction with constraints.*** From the dense matching, the software built a dense point-cloud-based 3D reconstruction and introduced geometric bounds to focus the reconstruction region on the edge-guardrail structures. This reduces

computational cost and improves reconstruction speed. The reconstruction result after guardrail identification is shown below.

Figure 23 Results of 3D reconstruction result after guardrail identification

5.3 Missing Guardrails Recognition and Localization

5.3.1 Simulated Design of Missing Guardrails

To simulate the potential absence of real-world guardrails as accurately as possible, the standard point cloud of the top guardrail area is divided into four regions based on orientation. The workflow for this area division is presented in Figure 24. Subsequently, a section of the guardrail is randomly removed from one of these regions to represent a typical 'missing guardrail' scenario.

Because the positions of UAV-collected images vary across different acquisition periods, the reference coordinate system of the model-derived from oblique photography and 3D reconstruction-also varies. Therefore, random rotation and translation transformations are applied to the coordinate system of the simulated scene

to replicate these reconstruction variations. A diagram illustrating the simulated missing guardrail scenarios and the random coordinate transformations is shown in Figure 25.

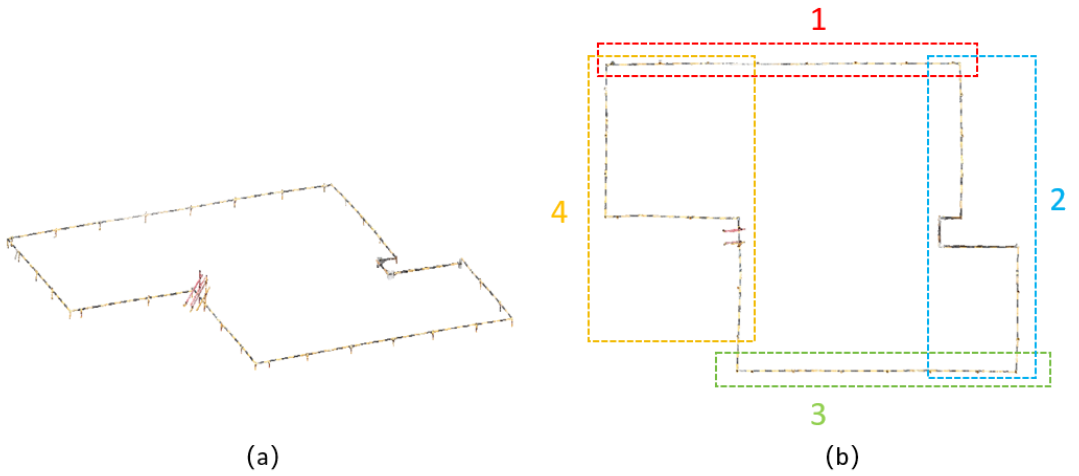


Figure 24 Standard Guardrail Point Cloud and Simulated Area Division: (a) standard point cloud of the top guardrail area; (b) the schematic diagram of area division

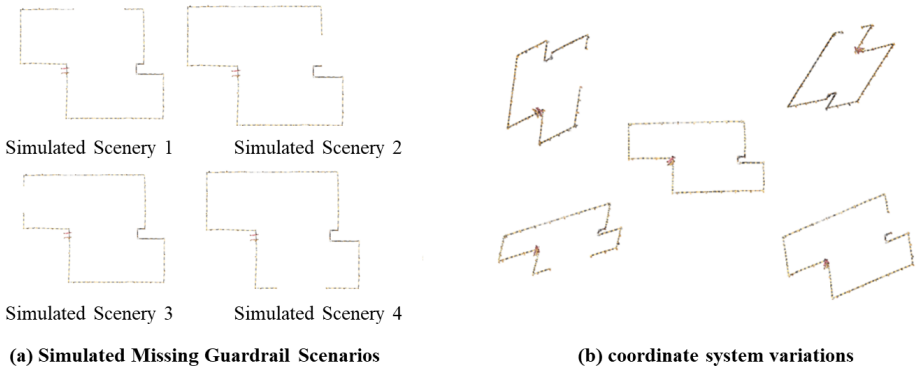


Figure 25. Four simulated missing guardrail scenarios and coordinate system variations

5.3.2 Voxel Filtering Down-sampling and Coarse Registration Algorithm

Point cloud registration typically requires point-pair feature matching and coordinate calculation between the source and target point clouds. However, high-resolution point cloud data captures not only macroscopic shapes but also microscopic geometric structures, leading to data redundancy and low computational efficiency. Downsampling is a widely recognized method for significantly reducing computational

load and optimizing registration time by decreasing point cloud density. Therefore, in this section, we propose a voxel filter downsampling method to effectively preserve the shape and detailed characteristics of the guardrail. A comparison between the initial point cloud and the voxel downsampled point cloud is shown in Figure 26.

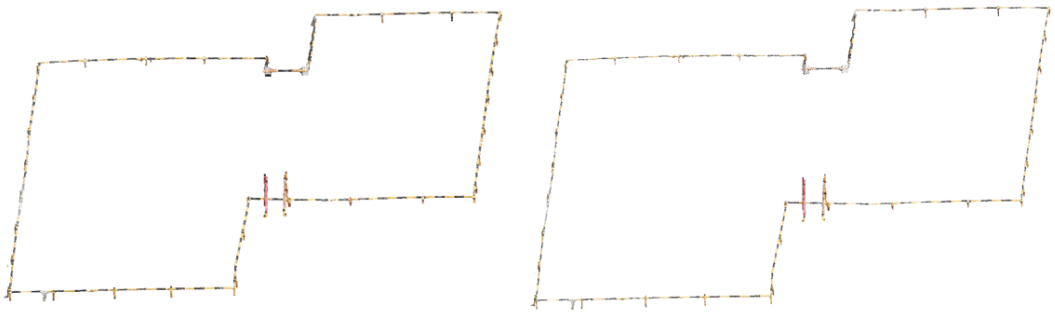


Figure 26 Comparison of the Initial point cloud and Downsample: (a) initial standard point cloud; (b) voxel downsampled point cloud

Since the initial pose of the standard guardrail point cloud differs significantly from that of the simulated missing guardrail point cloud, the direct ICP method (fine registration) fails to achieve precise registration after voxel downsampling. Therefore, the SAC-IA algorithm is applied to perform robust coarse registration, preventing the subsequent ICP algorithm from falling into local optima. Preliminary correspondences are established by calculating FPFH features, and RANSAC is used to filter out point pairs containing significant mismatches. Finally, the transformation supported by the maximum number of inliers is selected as the globally optimal initial pose. Figure 27

presents the comparison between the standard and simulated missing guardrail point clouds before and after coarse registration. The main steps of the algorithm are as follows:

1) Random matching of sampling points: according to FPFH features, n (n≥3) points are randomly selected from the source point cloud P, and their corresponding points in the target point cloud Q are determined by nearest neighbor search.

2) Estimating the initial transformation: compute an initial hypothetical transformation T using the selected sample matches n

3) Transform: apply this initial transformation to the entire source point cloud P

4) Interior point screening: the nearest neighbor search is carried out between the transformed source point cloud P and the target point cloud Q based on the Euclidean Distance Threshold, and then the interior point set satisfying the rule is screened out (repeat step 1 if the number of interior points is insufficient)

5) Optimize the transformation: use the new interior point pair relation to re-estimate a more accurate hypothetical transformation $\hat{T}$

6) Update the optimal result: calculate the distance between the corresponding points of the current interior point $\varepsilon(T)$. If the distance reaches a minimum, T is taken as the current optimal transformation; otherwise, repeat the above process until the set number of iterations is reached or the error converges.

$$\hat{T} = \arg min \mathcal{E}(T) = \arg min \sum_{p \in P} (Tp - q)^2 \qquad \text{Eq. [20]}$$

where p refers to the points of source point cloud P and q refers to the points in target point cloud Q.

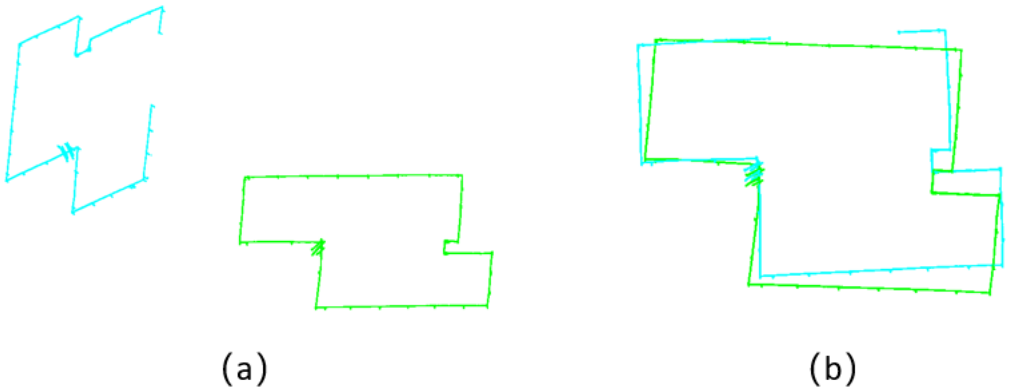


Figure 27 Comparison results of the standard guardrail point cloud and simulated missing guardrail point cloud in coarse registration process: (a) before coarse registration; (b) after coarse registration

5.3.3 Fine Registration Algorithm for Missing Guardrail Recognition

Although the simulated point cloud is coarsely aligned with the standard one via SAC-IA registration, KD-ICP fine registration is employed to further minimize alignment errors. KD-ICP accelerates the nearest neighbor search by constructing a KD-Tree, which significantly improves the efficiency of the ICP algorithm and demonstrates superior performance when processing large-scale scene point cloud data. Figure 28 illustrates the comparison between the standard and simulated missing guardrail point clouds before and after fine registration. The main process is summarized below:

1) Nearest Neighbor Search: Use the KD-Tree structure to quickly find nearest neighbor pairs in source and target point clouds

2) Rigid body transformation calculation: Compute the optimal rigid body transformation matrix by minimizing the error between the nearest neighbors of the source point cloud and the target point cloud

3) Use rigid body transformation: Apply the transformation to the source point cloud and update its position

4) Terminate transformation: Stop iteration when convergence condition satisfies minimum transformation error.

$$\varepsilon = \arg min K^{-1} \cdot \sum_{i=1}^{K} \| x_i - (M_R p_i + M_T) \|^2 \quad \text{Eq. [21]}$$

Where x_i and p_i represent the coordinates of the i-th point in point clouds X and P, respectively; M_R is the rotation matrix, and M_T is the translation vector of the rigid body transformation.

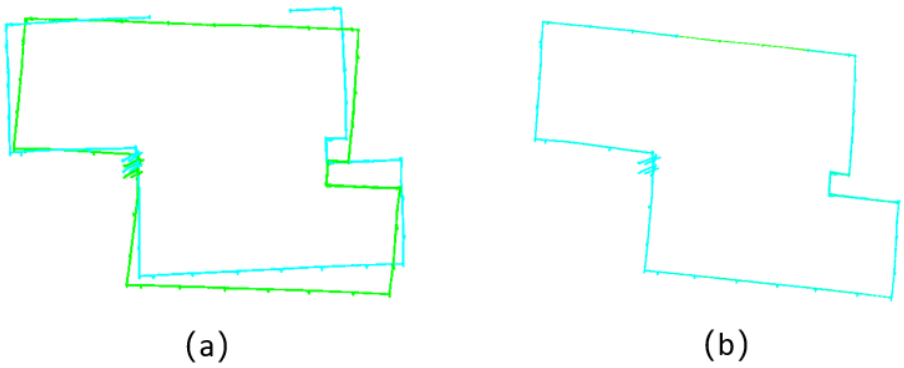


Figure 28 Comparison results of the standard guardrail point cloud and simulated missing guardrail point cloud in fine registration process: (a) before fine registration; (b) after fine registration

5.3.4 Nonoverlapped Detection for Missing Guardrail Recognition Localization

Following point cloud registration, it is necessary to detect the non-overlapping regions between the registered simulated point cloud and the standard guardrail point cloud. The difference between these clouds reveals the position of the missing guardrail. The main steps of the algorithm are as follows:

Step 1: Finding Overlapping Point Pairs

A nearest neighbor search is performed to establish correspondences between the source and target point clouds. Point pairs with distances below a predefined threshold are extracted, and their indices in the target point cloud are recorded to identify the overlapping regions.

Step 2: Removing Overlapping Regions

Based on the recorded indices, a radius search is conducted within a range of 0.05 m around each indexed point to fully define the overlapping point cloud. Subsequently, the points corresponding to this overlapping region are removed from the target point cloud. The remaining points constitute the non-overlapping point cloud, which represents the location of the missing guardrail. The detection results are shown in Figure 29.

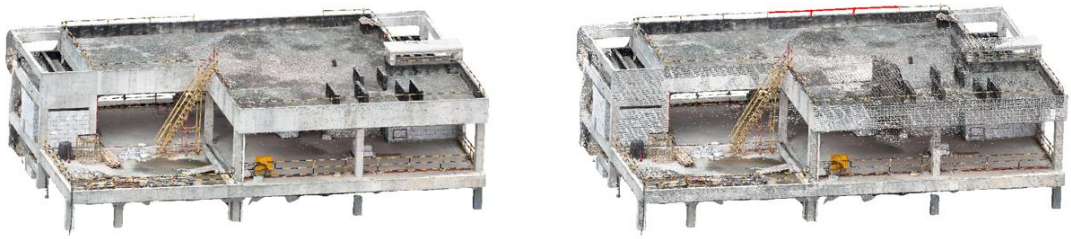


Figure 29 The detected result of the missing guardrail: (a) standard point cloud of simulated scenery 1 and (b) detected results of the missing guardrails in scenery.

5.4 Model Training and Implementation Details

Our experimental platform is built upon Ubuntu 20.04, equipped with an Intel Xeon Gold 6138 processor, 128 GB of system memory, an NVIDIA RTX A5000 GPU, PyTorch 2.0.1, CUDA 11.8, and Python 3.8.0. The batch size refers to 16, the input

image size to 448×448×3, and the number of epochs to 100. Stochastic gradient descent (SGD) serves as the optimizer within the model, with the initial learning rate and momentum set to 0.01 and 0.9, respectively.

For binary semantic segmentation with DeepLabv3+, we adopt the binary cross-entropy (BCE) loss; the formulation is presented by Eq. [22].

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad \text{Eq. [22]}$$

where y_i denotes the ground-truth label, $\hat{y}_i$ is the model prediction, and N is the number of samples.

5.5 Evaluation Performance Metrics

The evaluation on the detection performance of the improved segmentation model is defined by four key metrics: precision, Recall, mean Intersection over Union (mIoU), and F1 score. Among these, IoU represents the ratio of the intersection to the union between the predicted segmentation and the ground-truth segmentation, reflecting their degree of overlap—values closer to 1 imply better results. MIoU is the average IoU across all classes; larger values indicate better segmentation performance. The formulas for IoU and MIoU are as follows:

$$IoU = \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \quad \text{Eq.[23]}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \quad \text{Eq.[24]}$$

Where, k denotes the number of classes, $k + 1$ indicates that the background class is included, and i is the ground-truth class and j is the predicted class.

The F1 score is a metric for comprehensively evaluating a classification model's performance, regarding both precision and recall. Through the balance of the former two indicators, how well the model performs is directly reflected. The F1 score can be defined by Eq. [25]:

$$F_1 = 2 \cdot \frac{precision \cdot \text{recall}}{precision + \text{recall}} \quad \text{Eq. [25]}$$

5.6 Summary

This chapter develops an end-to-end pipeline that transforms 2D, pixel-level scene understanding into spatially anchored safety intelligence within a 3D reconstruction model. Starting from site UAV imagery, an improved DeepLabV3+ model is proposed to semantically segment safety guardrails and related context, producing per-pixel masks. These 2D results are then mapped onto an existing 3D reconstruction of the workspace via texture projection, yielding a semantically labeled mesh or point cloud in which each surface element inherits labels from one or more camera views. By grounding 2D predictions in the 3D reconstruction model, the computer vision system enables precise localization of unguarded edges and supports managers in making decisions regarding where guardrails are missing.

CHAPTER 6 AN IMPROVED VISION LEARNING APPROACH TO PREDICT HAZARDS IN CONSTRUCTION[2]

The occurrence of accidents is usually caused by two elements including unsafe conditions (objects) and unsafe behavior (people), based on the accident causation model. Once the unprotected edges (guardrails) have been clearly identified, unsafe behaviors like workers approaching unsafe construction areas should be further controlled. Existing studies tend to be retrospective, primarily examining the psychological factors contributing to unsafe behavior. Consequently, they often fail to visually capture the dynamic and complex conditions that influence these behaviors in real time. To address this limitation, this chapter presents a novel computer vision approach capable of automatically predicting potential unsafe behaviors—such as approaching hazardous areas with unguarded edges—by monitoring actions in real-time video. The proposed framework consists of three key components: (1) tracking individuals using SiamMask; (2) predicting pedestrian trajectories using an improved Social LSTM; and (3) identifying unsafe behaviors using Franklin's point-in-polygon algorithm (PNPoly)

Here, this research used SiamMask to track workers in real time to determine their on-site locations. Then, based on the tracked locations, we develop an improved Social Long Short-Term Memory (LSTM) (Alahi *et al*., 2016), a form of Recurrent Neural

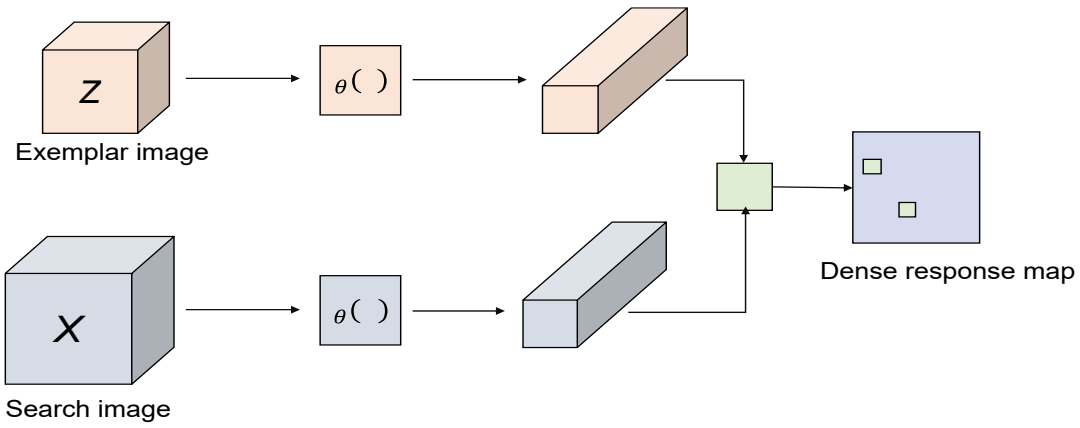
[2] This chapter is based on a published study.

Kong, T., Fang, W., Love, P. E.D, Luo, H., Xu, S., & Li, H. (2021). Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction. Advanced Engineering Informatics, 50, 101400.

Network (RNN), to learn their movement patterns. Once the model captures and learns these patterns, it can predict future trajectories. Finally, using Franklin's point-in-polygon (PNPoly) algorithm (Franklin, 2006), the spatial relationships between predicted trajectories and predefined hazardous zones are analyzed. This enables the inference and prediction of likely unsafe behaviors. This chapter focuses on the application of SiamMask, Social LSTM, and the PNPoly algorithm.

6.1 SiamMask-based Tracking

The SiamMask approach is adopted in this research to simultaneously achieve higher tracking performance and real-time detection speed (Perazzi *et al*., 2017). Chopra *et al*. (2005) first proposed the Siamese network for face recognition. A sample pair, consisting of an exemplar image *z* and a search image *x*, is input into the two branches of the network, which share the same weights. The Euclidean distance between their feature representations is then used to determine whether the exemplar and search images belong to the same category. SiamFC formulates the tracking problem as similarity matching learned in an embedding space. The structure of SiamFC is presented in Figure 30.



Adapted from Pinheiro *et al*., (2016)

Figure 30 Structure of a SiamFC

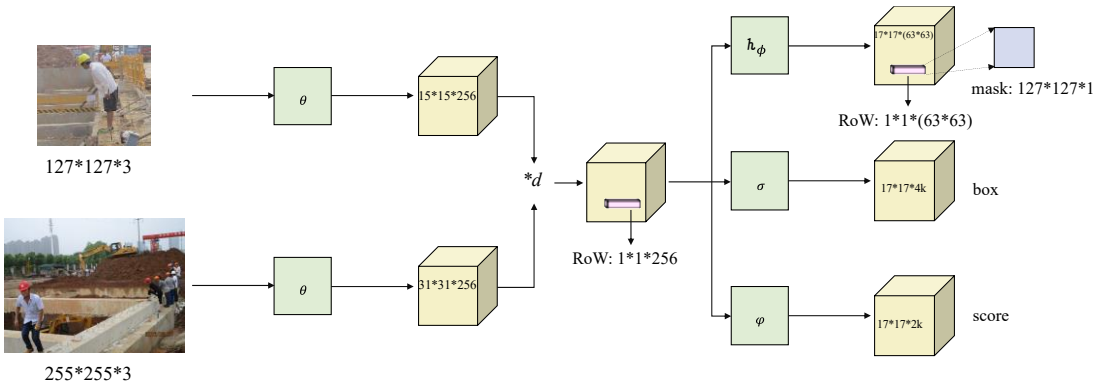
The SiamFC compares exemplar image *z* with search image *x* to obtain a dense response map. The two inputs are processed by the same networkf_θ, then output two feature maps. Finally, a cross-correlation between these two feature maps is performed, which generates a dense response map, as noted in Eq. [26].

$$g_\theta(z, x) = f_\theta(z) * f_\theta(x) \quad \text{Eq. [26]}$$

6.1.1 Network Architecture and Model Training

The structure of SiamMask is presented in Figure 24. A ResNet-50 network is used as a backbone network (f_θ) (Redmon *et al*., 2016). To obtain a high spatial resolution in deeper layers, we reduce the output stride to 8 using convolutions with stride 1. As noted in Figure 31, the exemplar and search images are inputs into the backbone network (f_θ). The outputs are separated feature maps (15*15*256 and 31*31*256). Then, these two feature maps are used to generate the response map through cross-correlation.

Here, we refer to each element of the response map as 'Response of a candidate Window' (ROW). Subsequently, the ROWs are sent into two 1×1 convolutional layers, one with 256 and the other with 63^2 channels, and output three-branch including mask, box generation, and score. The strategy of merging the low- and high-resolution features produces a more accurate object mask. The bounding box is generated from the obtained mask *via* a rotated minimum bounding rectangle (Ren *et al*., 2015; Perazzi *et al*., 2016).



Adapted from Wang *et al*., (2019)

Figure 31 Structure of SiamMask

6.2 Social-LSTM-based Trajectory Prediction

After tracking people in real time, the next step is to predict their trajectories. In this work, Social-LSTM is adopted due to its state-of-the-art performance on standard datasets such as ETH and UCY (Seo *et al*., 2016). By modeling person-person interactions during trajectory prediction, Social-LSTM improves both the robustness and accuracy of multi-person tracking.

The Kalman filter is employed to refine the predictions generated by the Social-LSTM model, thereby enhancing the overall robustness of the trajectory forecasting method. As an optimal recursive estimator, the Kalman filter infers relevant state parameters from noisy, indirect, and uncertain measurements, making it particularly suitable for estimating a person's movement based on their historical trajectory data. Accordingly, the Kalman filter is applied to adjust the Social-LSTM outputs whenever discrepancies arise between a person’s predicted walking speed and their actual recorded speed. The workflow of this improved Social-LSTM framework is illustrated in Figure 32.

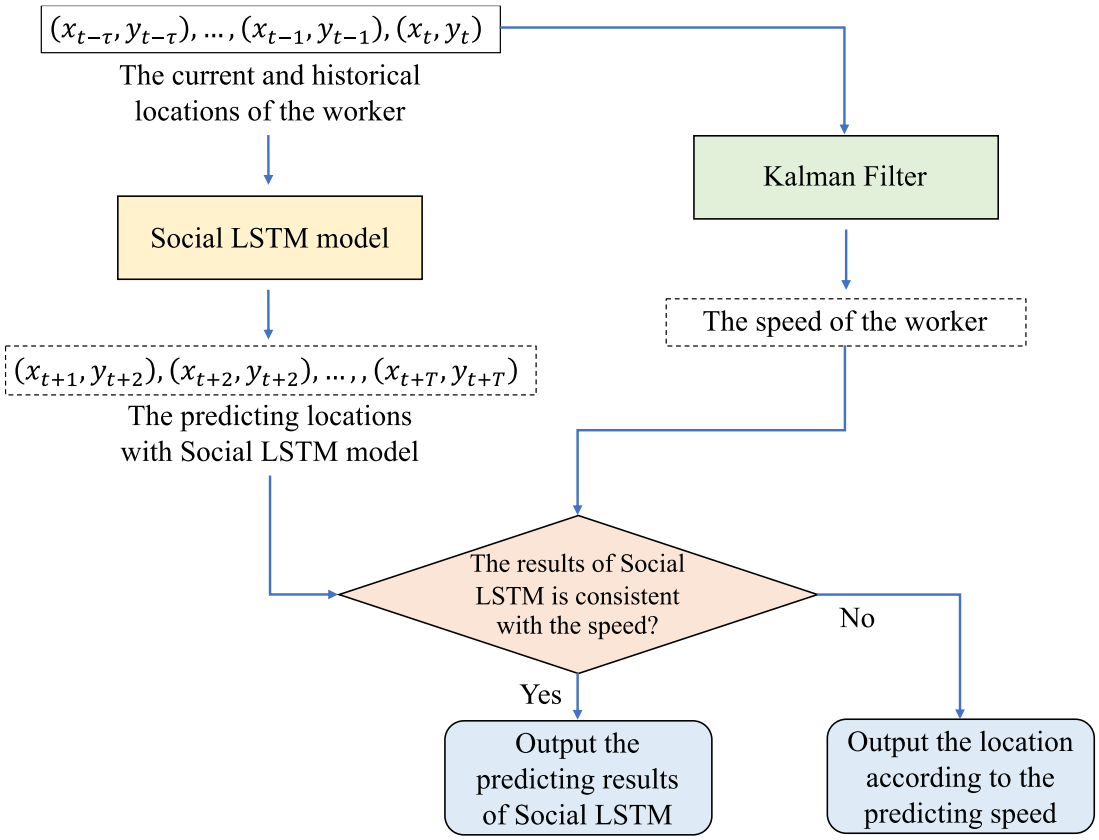


Figure 32 Process of improved Social LSTM

A social pooling layer is incorporated between time step t and time step t+1. The output is a three-dimensional state tensor, where the first two dimensions represent plane coordinates and the third dimension corresponds to the state tensor output by the Social-LSTM model at time step t. This structure enables the prediction of the target person's trajectory by aggregating Social-LSTM information from neighboring individuals. The procedure is as follows:

Step 1: Capture the hidden state information (h_i^t) of the i-th person at time-step t;

Step 2: Share hidden state information with neighbours by building a 'social' hidden-state tensor H_t^i, as noted in Eq. [27]:

$$H_t^i(m,n,:) = \sum_{j \in N_i} 1_{mn}\left[x_t^j - x_t^i, y_t^j - y_t^i\right] h_{t-1}^j \quad \text{Eq. [27]}$$

where, D is the hidden-state dimension; N_o is the neighbourhood size; h_{t-1}^j is the hidden state of the Social-LSTM corresponding to the j^{th} person at time-step t-1; $1_{mn}[x,y]$ is an indicator function to check if (x,y) is in the (m,n) cell of the grid; N_i is the set of neighbours corresponding to the person i.

Step 3: Embedding the pooled social hidden-state tensor into the vector a_t^i and embedding coordinates into vector e_t^i, as noted in Eq. [28], Eq. [29] and Eq. [30].

$$h_i^t = LSTM(h_i^{t-1}, e_i^t, a_t^i, W_l) \quad \text{Eq. [28]}$$

$$a_t^i = \phi(H_t^i; W_a) \quad \text{Eq. [29]}$$

$$e_t^i = \phi(x_t^i, y_t^i; W_e) \quad \text{Eq. [30]}$$

where, $\phi(\cdot)$ is an embedding function with ReLU non-linearity; W_e and W_a embedding weights; W_l is the Social-LSTM weight.

6.2.1 Trajectory Estimation

The hidden-state information at time-step t is used to predict the distribution of the trajectory position $(\hat{x}, \hat{y})_{t+1}^i$ at the next time-step t + 1. Following Graves (2013), a

bivariate Gaussian distribution parametrized is assumed with the mean $\mu_{t+1}^{i} = (\mu_x, \mu_y)_{t+1}^{i}$, standard deviation $\sigma_{t+1}^{i} = (\sigma_x, \sigma_y)_{t+1}^{i}$ and correlation coefficient ρ_{t+1}^{i}. The predicted coordinates $(\hat{x}_t^i, \hat{y}_t^i)$ at time-step *t* can be computed by flowing Equations.

$$(\hat{x}, \hat{y})_t^i \sim N(\mu_t^i, \sigma_t^i, \rho_t^i) \quad \text{Eq. [31]}$$

$$[\mu_t^i, \sigma_t^i, \rho_t^i] = W_p h_i^{t-1} \quad \text{Eq. [32]}$$

$$L^i(W_e, W_l, W_p) = -\sum_{t=T_{obs}+1}^{T_{pred}} log(\mathbb{P}(x_t^i, y_t^i | \sigma_t^i, \mu_t^i, \rho_t^i)) \quad \text{Eq. [33]}$$

where, L^i is the i^{th} trajectory.

6.3 Prediction of Unsafe Behaviour

After predicting a person's future trajectory, a PNPoly algorithm determines if their coordinates are in the hazardous area so that unsafe behaviour can be envisaged. As shown in Figure 33, if the predicted trajectories (P_m and P_k) are within the polygon, then a person's behaviour will be envisaged as unsafe.

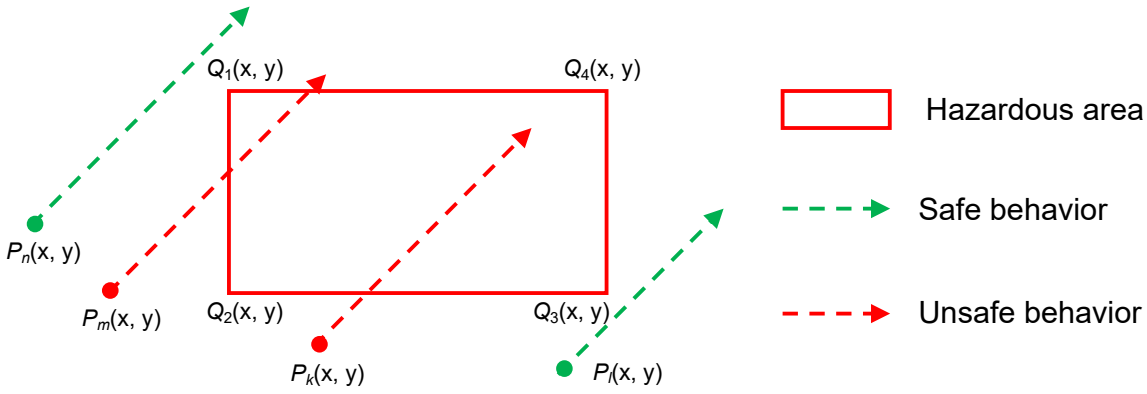


Figure 33. Example of safe and unsafe behaviour based on the predicted trajectory

The PNPoly algorithm tests whether a point is inside a polygon (convex or concave) by counting how many times the ray from the test point crosses its edge. If the count is an odd number, the point is in the polygon area; otherwise, it is outside.

6.4 Model Training and Implementation Details

The output of the SiamMask model includes mask, bounding box, and score. Hence, the training loss of SiamMask contains three parts, which the following equations can compute.

$$L_{loss} = \lambda_1 L_{mask} + \lambda_2 L_{score} + \lambda_3 L_{box} \quad \text{Eq. [34]}$$

$$L_{mask} = \sum_n \left(\frac{1+y_n}{2wh} \sum_{ij} log\left(1 + e^{-c_n^{ij} m_n^{ij}}\right) \right) \quad \text{Eq. [35]}$$

$$L_{score} = -logp_x \quad \text{Eq. [36]}$$

$$L_{box} = \sum_{i=0}^{3} smooth_{L_1} = \sum_{i=0}^{3} smooth_{L_1}(\delta[i], \sigma) \quad \text{Eq. [37]}$$

$$smooth_{L_1}(x, \sigma) = \begin{cases} 0.5\sigma^2 x^2 & |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2} & |x| \geq \frac{1}{\sigma^2} \end{cases} \quad \text{Eq. [38]}$$

where, λ_1, λ_2, and λ_3 are hyper-parameters, set to 32, 1 and 1, respectively; y_n is the ground-truth binary label of each ROW, wherey_n is set to 1 following Wang *et al.*(2019); c_n^{ij} The pixel-wise ground-truth label corresponds to a pixel (*i*, *j*) of the

object mask in the n-th candidate ROW; w and h are the sizes of a pixel-wise ground-truth mask.

6.5 Evaluation Performance Metrics

This section evaluates object tracking and trajectory prediction using two families of metrics: tracking metrics and trajectory-prediction metrics.

(1) Evaluation performance metrics for object tracking

Three key performance indicators (KPIs) are used to evaluate the tracking performance of our approach: (1) mean intersection over union (mIOU); (2) mean average precision (AP)@0.5 IOU; and (3) mean AP@0.7 IOU. The AP is the area under the precision and recall curve. Here, the IOU, precision and recall can be computed in Eq. [39], Eq. [40], and Eq. [41]:

$$IOU = \frac{Detection\ result \cap Ground\ truth}{Detection\ result \cup Ground\ truth} \quad \text{Eq. [39]}$$

$$precision = \frac{TP}{TP+FP} \quad \text{Eq. [40]}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{Eq. [41]}$$

where, A 'true positive' (TP) refers to a person who has been correctly tracked. A 'false positive (FP)' occurs when a tracked worker is some other object. A 'false negative (FN)' refers to a failure in tracking a person in the image.

(2) Evaluation performance metrics for trajectory prediction

Two key performance indictors (KPIs) are used to evaluate the trajectory prediction performance: (1) average displacement error (ADE); and (2) final displacement error (FDE). These two KPIs can be computed in Eq. [42] and Eq. [43]:

$$ADE = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{t_{pred}}\sum_{t=t_{obs}+1}^{t_{obs}+t_{pred}}\sqrt{(x_i^t - \hat{x}_i^t)^2 + (y_i^t - \hat{y}_i^t)^2} \quad \text{Eq. [42]}$$

$$FED = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_i^{t_{pred}} - \hat{x}_i^{t_{pred}})^2 + (y_i^{t_{pred}} - \hat{y}_i^{t_{pred}})^2} \quad \text{Eq. [43]}$$

A video from our real-time monitoring system is selected and used to validate the performance of Social-LSTM for trajectory prediction. In this experiment, the hyperparameter of the Social-LSTM is set as follows: the spatial pooling size (N_0) is 32, the pooling window size is 8*8, and the learning rate is 0.003.

6.6 Summary

This chapter presents a real-time, computer vision-based approach to proactively predict unsafe worker behavior on construction sites by forecasting whether a worker will approach or enter hazardous zones without edge protection. Our proposed approach consists of three components: (1) tracking workers with SiamMask to localize them on site; (2) modeling and predicting worker motion using an improved Social-LSTM; and (3) checking predicted positions against mapped hazard polygons via Franklin's

PNPoly point-in-polygon test. By fusing tracking, trajectory forecasting, and spatial inclusion analysis, the proposed computer vision system can infer likely unsafe actions before they occur, enabling early intervention and enhanced on-site safety.

CHAPTER 7 EXPERIMENTS AND RESULTS

This chapter targeted to systematically evaluate the proposed approaches and demonstrate their effectiveness and feasibility. We describe the dataset development and model implementation details, define evaluation metrics, and conduct ablation studies. Key research findings and practical implications are summarized at the end.

7.1 Results and Analysis of Missing Guardrails Recognition

To evaluate the effectiveness and feasibility of the proposed approach, a real-life high-rise building project in Wuhan, China, is used as a case study. This project serves as a flagship development in the East Lake High-tech Development Zone—also known as the Optics Valley of China (OVC). The Wuhan New City concentrates major public buildings, research facilities, and headquarters projects within a 719-km² strategic area prioritized in provincial and municipal initiatives for accelerated delivery.

The first construction phase of the project has a planned gross floor area of approximately 179,400 m². Within this phase, the initial southwest sub-area, comprising the FinTech Service Center, the Investor Tower, and affiliated buildings with covers roughly 90,000 m². Project mobilization commenced on 17 November 2023, marking its inclusion among the first wave of key projects to enter full execution. By May 2025, the main structures of the initial sub-area had been topped out, reflecting rapid on-site progress. The overall completion of Phase I is targeted for December 2025, as indicated in earlier official communications.

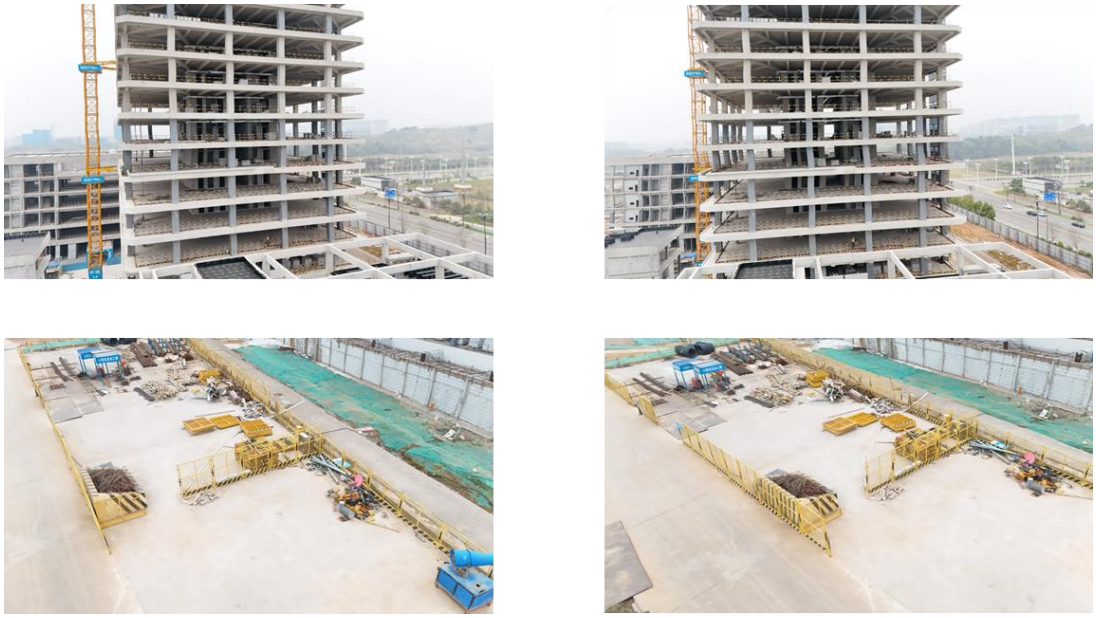


Figure 34 Example of case project

7.1.1 Data Collection and Processing

We used a DJI Matrice 300 RTK (M300 RTK) equipped with an H20T hybrid camera to acquire high-resolution RGB imagery of edge protection (guardrails) at the construction site. Real-time kinematic (RTK) positioning was recorded concurrently to ensure centimeter-level georegistration. Flight routes were pre-planned in DJI Pilot with terrain-following lines offset 5-10 m from the slab edge. At a flight altitude of 20 m, the mission targeted a ground sampling distance (GSD) ≤ 5 mm/pixel. To mitigate occlusions, we flew a cross-grid pattern combining nadir and oblique views and limited each mission's coverage to approximately 200 m × 200 m. Key platform specifications are summarized in Table 2.

Table 2 Parameters of the DJI M300

Parameter	Value
Takeoff weight	720 g

Max flight time	46 min
Max wind resistance (wind speed)	12 m/s
Hovering accuracy	Vertical: ±0.1 m; Horizontal: ±0.3 m
Max ascent speed	10 m/s
Max descent speed	10 m/s
Max horizontal flight speed	21 m/s
Max tilt angle	35°
Effective camera resolution	48 MP

The experimental dataset comprises images of edge guardrails on multiple floor platforms at the CSCEC Third Bureau Sci-Tech Finance Headquarters, captured using a DJI Air 3 drone. A total of 551 source images were collected at an original resolution of 4032 × 2268 pixels from diverse viewpoints. Each image was resized and randomly cropped to 448 × 448 pixels, resulting in a final working dataset of 1,728 images.

To enhance model robustness, we applied four augmentation techniques—horizontal flip, in-plane rotation about the image center, additional random crops, and saturation adjustment—to expand the dataset threefold, resulting in 5,184 images. Based on the augmented datasets, 4,784 images are used for training, 200 images for validation, and 200 images for testing, separately.

7.1.2 Edge Guardrail Detection

To comprehensively verify the performance gains contributed by each improved module, we conducted ablation experiments on the proposed model. The experiments were built upon YOLOv11 and employed a stepwise integration strategy to quantify the contribution of each module to detection accuracy and efficiency. We evaluated three comparative configurations: the pre-optimization YOLOv11 (Baseline), YOLOv11 with FRFN, and YOLOv11 with SAFM. Table 3 exhibited the ablation results, where the Baseline denotes the vanilla YOLOv11 model.

Table 3 Ablation results

Baseline	FRFN	SAFM	Precision	Recall	mAP0.5	mAP0.5-0.95
√			0.7936	0.7094	0.7742	0.5177
√	√		0.8476	0.7547	0.8272	0.5567
√		√	0.8491	0.7520	0.8274	0.5503
√	√	√	0.8556	0.7607	0.8305	0.5576

Ablation experiments indicate that both FRFN and SAFM contribute substantial performance gains. Their combination (YOLOv11-FS) attains the highest scores on all four metrics—Precision, Recall, mAP@0.5, and mAP @[0.5:0.95]. As shown in Figure 35, the training-result comparison between YOLOv11-FS and the YOLOv11 further confirms the effectiveness of the proposed network for guardrails detection.

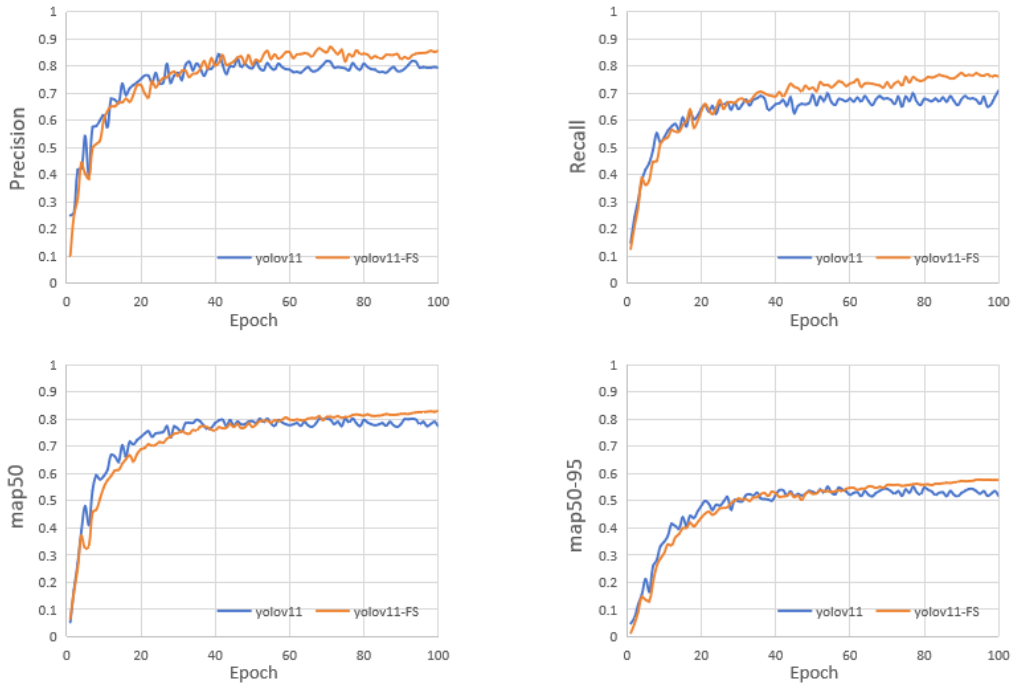
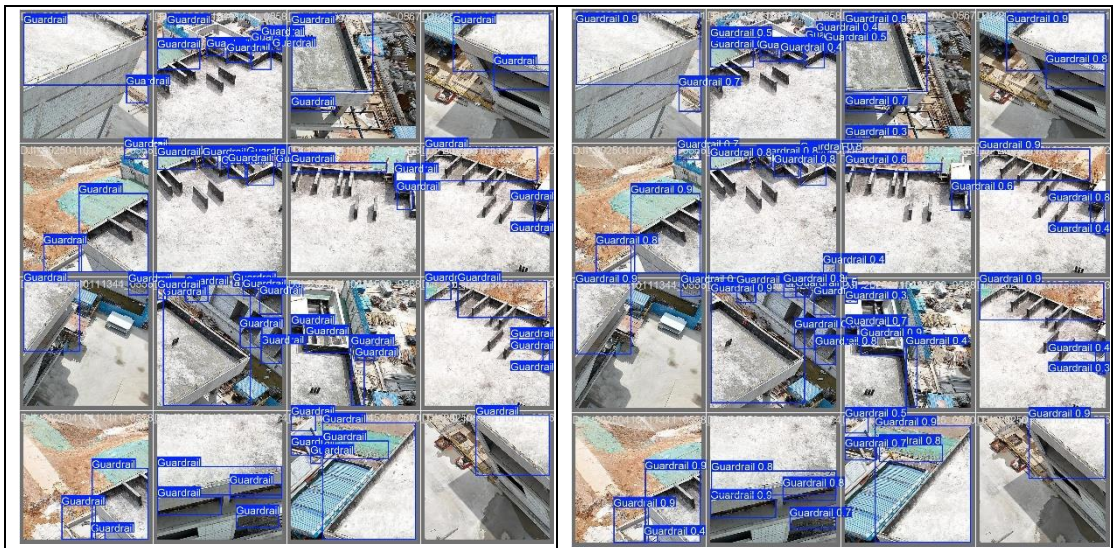


Figure 35. Comparison of results between YOLOv11-FS and YOLOv11.

We randomly selected twelve images from the test set and performed detection with YOLOv11-FS. Figure 36(a)-(b) juxtaposes human annotations with the model's predictions. Across varying viewpoints and substantial scale differences of the guardrails, the model maintains accurate detections, demonstrating strong robustness.



(a) (b)

Figure 36 YOLOv11-FS detections and annotations (ground truth).

7.1.3 Edge Guardrail Segmentation

To comprehensively assess the performance gains brought by each enhanced module, a series of ablation experiments has been performed our proposed model using the above dataset. Built upon DeepLabv3+, the study adopts a stepwise integration strategy to quantify each module's contribution to accuracy and efficiency. Three comparative configurations were evaluated: the pre-optimization DeepLabv3+ (Baseline), DeepLabv3+ with the SE module, and DeepLabv3+ with the CBAM module. The ablation results are summarized in Table 4.

Table 4 Ablation results

Baseline	SE	CBAM	F1	Precision	Recall	Miou
√			0.8397	0.8643	0.8165	0.7509
√	√		0.8663	0.8864	0.8470	0.7842
√		√	0.8555	0.8796	0.8326	0.7694
√	√	√	0.8717	0.8928	0.8528	0.7936

Experimental results indicate that each improved module has a significant effect on enhancing crack image segmentation performance. Specifically, when only the SE block is added, the F1 score, precision, recall, and mIoU increase by 2.66%, 2.21%, 3.05%, and 3.33%, respectively, compared with the baseline, indicating that the SE block strengthens the extraction of features for edge guardrails and improves the model's predictive performance. When only the CBAM block is added, the F1 score, precision, recall, and mIoU increase by 1.58%, 1.53%, 1.61%, and 1.85%, respectively,

relative to the baseline. This suggests that the CBAM block enhances the model's ability to adaptively reweight feature importance, enabling it to focus more on salient features while suppressing less important ones, thereby improving feature-sensitivity in perception. When the SE and CBAM blocks are integrated, the F1 score, precision, recall, and mIoU improve by 3.20%, 2.85%, 3.63%, and 4.27%, respectively, over the baseline. These results provide strong evidence for the effectiveness of the improved model. Figure 37 present an example result of Deeplabv3Plus-SC and Deeplabv3Plus.

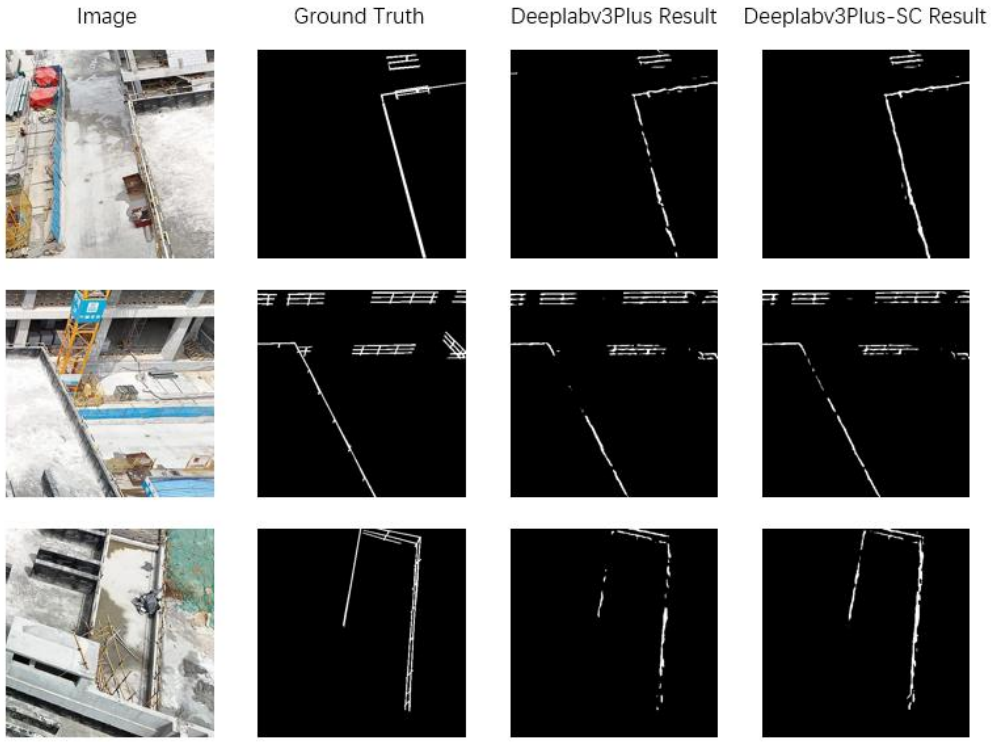


Figure 37 Example results of Deeplabv3Plus-SC and Deeplabv3Plus.

As shown in Figure 37, we compare the improved Deeplabv3Plus-SC model with the original Deeplabv3Plus and the ground-truth segmentation. Visually, the Deeplabv3Plus-SC results are closer to the ground-truth segmentation than those of the original model. This advantage is especially evident in complex scenes where the guardrails have low contrast with the background: the improved model yields more

complete guardrail segmentations, indicating greater robustness. These gains support better performance in subsequent guardrail texture mapping and 3D reconstruction.

7.1.4 Grad-CAM Visualization

To interpret the model's decision process for guardrail detection, we visualize discriminative regions using Grad-CAM. The pipeline is as follows. We first perform a forward pass through the backbone and neck, and obtain class scores from the classification branch of the detection head. We then back-propagate the gradients of the target class score with respect to the convolutional feature maps in the head; these gradients indicate each map's contribution to the final prediction. Next, we apply global average pooling to the gradients to derive channel-wise weights. The class activation map is produced by a weighted sum of the feature maps and is passed through a ReLU to retain responses that positively influence the target class. Finally, we normalize and upsample the map to the input resolution and overlay it on the original image to obtain the Grad-CAM heatmap. The following section presents qualitative analyses of the visualizations over guardrail regions.

We selected guardrail images from multiple viewpoints and applied Grad-CAM to the three detection heads (small/medium/large scale) of YOLOv11-FS at their output layers. The visualizations reveal systematic differences in attention across heads driven by the object scale in the image. For near-view, large-scale guardrails, the small- and medium-scale heads exhibit strong responses along guardrail edges, whereas the large-scale head

primarily activates over the ground enclosed by the guardrail. Conversely, for distant, small-scale guardrails, the small- and medium-scale heads effectively capture edge information, while the large-scale head shows negligible activation—indicating that the small/medium heads are critical for detecting small guardrail targets. These Grad-CAM analyses clarify the model's attention patterns during guardrail detection and, in turn, enhance interpretability.

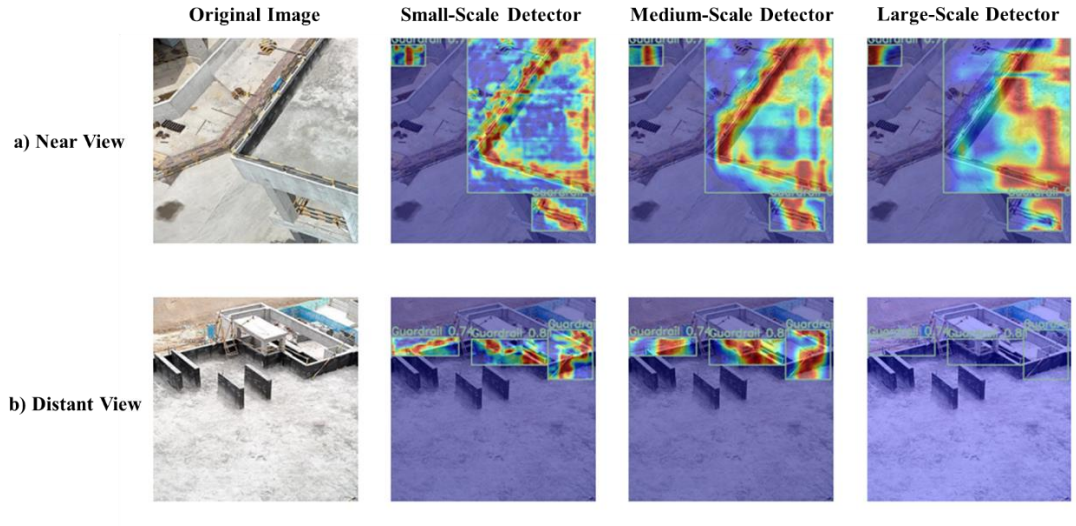


Figure 38 Explainable Visualization Results for Multi-View Images

7.1.5 Missing Guardrails Recognition

In this experiment, registration accuracy is evaluated quantitatively by RMSE (root mean square error) of point cloud registration, the smaller RMSE is, the higher registration accuracy is.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \| P_i - Q_i \|^2} \qquad \text{Eq. [44]}$$

From Table 5, it can be seen that the RMSE with the lowest accuracy is 6.276mm and the total registration time is 770.64s. For the missing guardrail recognition, the detection is generally done in centimeters, and the registration accuracy can meet the accuracy requirements of the missing guardrail recognition.

Table 5 Registration accuracy and runtime of four simulation scenarios

Simulation Scenarios	RMSE/mm	Coarse Registration/s	Fine Registration /s	Time/s
1	6.261	734.59	25.50	760.09
2	6.276	726.86	22.42	749.28
3	6.178	743.95	26.69	770.64
4	6.149	731.47	26.71	758.18

7.2 Results and Analysis of Unsafe Behavior Prediction[3]

The Wuhan Metro project is selected and used in this experiment to evaluate the feasibility and effectiveness of the proposed approach. This research specifically focusses on foundation pits and predicts people approaching their edge. Here, several Close-Circuit Television (CCTV) cameras (i.e., DS-2DF8223IW-A) with two million pixel high-definition were installed on construction sites and used to record workers' daily activities in real-time.

7.2.1 Tracking

The SiamMask model was pre-trained on the DAVIS database (Perazzi *et al.*, 2016). Then a video from a real-time monitoring system is collected for testing. The duration of this video is 222 seconds, including 5550 frames and includes four people. Figure 6 presents an example of people being tracked using the SiamMask approach. The mIOU,

[3] This section is based on a published study.

Kong, T., Fang, W., Love, P. E.D, Luo, H., Xu, S., & Li, H. (2021). Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction. Advanced Engineering Informatics, 50, 101400.

mAP@0.5, and mAP@0.7 of tracking these four people are 73.4%, 92.9%, and 68.1%, respectively. The testing speed of tracking each person is 0.018s.

The comparative analysis are conducted on two representative methods, the ASLA (visual tracking via Adaptive Structural Local Sparse Appearance Model) (Jia *et al*., 2012) and SCM (tracking via a Sparse Collaborative appearance Model) (Zhong *et al*., 2014), to judge the tracking performance of our model since they have been proved to extend high accuracy in related tasks (Xiao & Zhu, 2018). The average sequence overlaps score (AOS) and tracking length (TL) in Xiao and Zhu (2018)'s work for ASLA and SCM were 84%, 72%, and 83%, 68%, respectively. In this comparison study, we use the same testing database, and the parameters of ASLA and SCM accord with Jia *et al*. (2012) and Zhong *et al*. (2014), respectively. The results of the comparison study are presented in Table 6, where it can be seen that our adopted SiamMask approach improves our ability to track people on construction sites.

Table 6. Comparison of results to track worker

Methods	mIoU					mAP@0.5					mAP@0.7				
	#1*	#2	#3	#4	mean	#1	#2	#3	#4	mean	#1	#2	#3	#4	mean
ASLA	0.453	0.061	0.649	0.621	0.399	0.523	0.055	0.924	0.914	0.545	0.233	0.031	0.328	0.712	0.184
SCM	0.275	0.102	0.556	0.68	0.368	0.344	0.087	0.653	0.864	0.431	0.062	0.08	0.173	0.553	0.174
SiamMask	0.734	0.682	0.779	0.755	**0.734**	0.906	0.873	0.987	0.942	**0.929**	0.677	0.550	0.797	0.731	**0.681**

*Note: #1, #2, #3, and #4 are the number of people in our video; “mean” is the average result of these four people.

7.2.2 Trajectory Prediction

The ADE and FED of our model to 1s, 2s, 3s, 4s, 5s, 6s, 7s, 8s, 9s, and 10s are used to examine the performance of our model to predict future trajectory. The results presented in Figure 39 indicate that our improved Social-LSTM can achieve higher accuracy (e.g., ADE and FDE) on trajectory prediction except at the length of 1s. The ADE and FDE for our improved method at 1s were 7.019 pixel and 9.696 pixels, respectively. While the ADE and FDE for the Social-LSTM at 1s were 6.303 pixels and 11.505 pixels. Therefore, we can suggest that our improved Social-LSTM achieved robust results for long-term prediction. The testing speed of prediction of each person's trajectory is 0.030s.

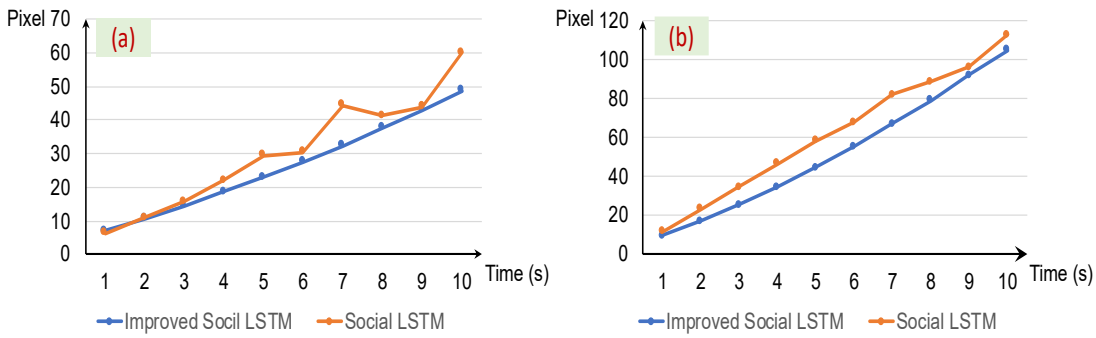


Figure 39. Results of trajectory prediction: (a) ADE; (b) FDE

Two deep learning models, the Social-LSTM and GANS are used to examine the tracking performance Gupta *et al*., (2018). The Social-GAN model has been previously used by Kim *et al*. (2019) to predict people's paths. The hyper-parameters of the Social-LSTM and GANs used in this comparison study were followed by Alahi *et al.* (2016) and Gupta *et al*. (2018), respectively. The results of these three models are presented in Table 7, and we can conclude that our improved Social-LSTM method outperforms

both the Social-LSTM and GANs.

Table 7. Result of comparison study on trajectory prediction

Model	**1.6s**		**2.4s**	
	ADE (pixel)	FDE (pixel)	ADE (pixel)	FDE (pixel)
Social LSTM+Kalman filtering	9.165	**14.161**	**12.132**	**20.470**
Social-LSTM	**9.014**	18.535	12.739	27.740
Social-GAN	12.23	19.64	16.15	27.35

Figure 40 present an example of the trajectory prediction by using these three methods. Here it can be seen that the: (1) blue points are the ground truth; (2) yellow points are predicted by using our improved Social-LSTM; (3) the cyan points are predicted by the Social-LSTM; and (4) red points are predicted by the Social-GAN.

(a) T=frame 55 (b) T=frame 755

Figure 40. Example of people trajectory prediction results

7.2.3 Unsafe Behaviour Prediction Results

A testing database is used to evaluate the effectiveness and feasibility of our approach that was used to predict unsafe behaviour. The hazardous area is manually defined and

labelled, as noted in Figure 34. An example of unsafe behaviour prediction is presented in Figure 41. Based on the people's predicted trajectories, we observe that two people (ID2 and ID3) have entered a hazardous area, while person #4 is predicted to arrive. Notably, our approach fails to predict unsafe behaviour as their trajectories are not correctly determined. Possible reasons influencing the performance of a person's trajectory are occlusions and their distance from the video camera.

Figure 41. Examples of unsafe behaviour prediction result

CHAPTER 8 CONCLUSIONS AND DISCUSSION

8.1 Conclusion

Falls from height (FFH) are consistently recognized as a leading cause of injuries and fatalities in the construction industry, with the absence or inadequacy of safety guardrails being a major contributor to FFH risk. Typically, manual inspections are labor-intensive and time-consuming, hindering their practical application on construction sites due to the complex and dynamic nature of the working environment.

Emerging technologies, such as Building Information Modeling (BIM) and Computer Vision (CV), have been extensively utilized for monitoring FFH risks. However, several limitations persist that hinder their deployment on construction sites:

1. Existing deep learning-based AI monitoring systems often employ complex, non-linear architectures that lack transparency. This 'black box' nature makes it difficult for on-site managers to interpret and trust the system's output.
2. Previous studies primarily rely on 2D image-based approaches or standalone 3D reconstruction for inspection. These methods often fail to precisely localize hazards within as-built 3D models or associate them with specific building components.
3. Many studies are retrospective, focusing on the psychological factors contributing to unsafe behavior. Consequently, they lack the ability to visually capture the dynamic, complex on-site conditions that directly influence unsafe actions in real time.

To address these challenges, a Design Science Research (DSR) paradigm is adopted to design, implement, and evaluate an XAI-enabled, UAV-based computer vision system. This system aims to: (1) detect unprotected edges and guardrails; (2) precisely localize FFH hazards within a three-dimensional (3D) construction site model and link them to specific building components; and (3) predict workers approaching unprotected edges based on video analysis. The main findings of our research are as follows

1) We designed a novel multi-scale detector based on YOLOv11, integrating FRFN and SAFM modules to enhance feature extraction and representation. Ablation studies were conducted to evaluate the individual contribution of each module, demonstrating consistent improvements over the baseline YOLOv11 across standard detection metrics—including precision, recall, mAP@0.5, and mAP@[0.5:0.95]. Experimental results indicate that the proposed approach achieves a precision of 0.8556, recall of 0.7607, mAP@0.5 of 0.8305, and mAP@[0.5:0.95] of 0.5576.
2) We addressed the YOLOv11-FS model interpretability through Grad-CAM visualizations. The generated heatmaps reveal coherent attention patterns tied to object scale. The experiment findings demonstrated that both explain the empirical gains from our multi-scale design and increase practitioner confidence in failure analysis and model debugging.
3) We propose an enhanced DeepLabv3+ network for semantic segmentation of

safety guardrails in site imagery. The resulting masks are projected onto a 3D reconstruction via texture mapping, yielding a semantically labelled mesh or point cloud. By anchoring 2D predictions in 3D, the method enables precise localisation of unguarded edges. Overall, the approach achieves F1 = 0.8717, precision = 0.8928, recall = 0.8528, and mIoU = 0.7936.

4) This research proposed an improved computer vision approach to predict worker approaching hazardous area. Workers are tracked using SiamMask, and an improved Social-LSTM predicts future trajectories. Hazard proximity is determined via Franklin's PNPoly point-in-polygon test against predefined danger zones. On a four-subject evaluation, the tracker achieves mIoU = 73.4%, mAP@0.5 = 92.9%, and mAP@0.7 = 68.1%, with a per-person inference time of 0.018 s. The forecaster attains ADE@1 s = 6.303 px and FDE@1 s = 11.505 px.

8.2 Limitations

Despite the contributions of this research, several limitations should be acknowledged and addressed in future work. First, domain shift caused by variations in illumination, weather conditions, and camera characteristics may degrade detection performance when applied beyond the current dataset. Since our evaluation is based on a single case study, differences in project scale, operational conditions, and environmental factors are likely to affect model accuracy. Future research will explore domain adaptation and transfer learning strategies to improve generalization across diverse settings.

Second, heavy occlusions caused by equipment or protective netting remain a significant challenge, adversely affecting both detection performance and the completeness of 3D reconstructions. Occluded regions lead to missing observations, which reduce the fidelity of the georeferenced site model.

Third, the volume of training data and the presence of label noise limit achievable accuracy, particularly for rare or atypical guardrail configurations. We plan to address this by curating a larger, multi-site dataset and implementing stricter annotation protocols and quality assurance measures to minimize label inaccuracies.

Finally, while Grad-CAM provides qualitative visual explanations, it does not quantify predictive uncertainty—a critical aspect for safety-sensitive decision-making. Future work will integrate uncertainty quantification and calibration methods (e.g., aleatoric and epistemic uncertainty estimates) to enhance model robustness and facilitate risk-aware interventions.

8.3 Suggestions on Future Research Directions

To further enhance the robustness and practical feasibility of the proposed monitoring system, future work may proceed along the following research directions:

1) Future research should focus on enhancing the system's robustness under complex site conditions. First, RGB imagery should be augmented with depth or thermal modalities to reduce false positives and negatives in low-illumination, strong

backlight, and occluded scenes. Additionally, multi-view fusion and next-best-view flight planning will be explored to mitigate occlusion. Second, the approach will be extended from still images to video through temporal modeling and object tracking, thereby stabilizing cross-frame predictions and suppressing jitter. These improvements are expected to enable more actionable compliance checks while maintaining a balance between sensor payload and computational complexity. Suitable evaluation metrics will include stratified recall and mAP across different scene conditions, temporal consistency measures, and 3D geometric deviation.

2) To address the challenge of detecting small, distant guardrails, future work will systematically investigate multi-scale representation mechanisms, including efficient feature pyramids, feature-enhancement modules, and attention mechanisms. Furthermore, domain adaptation and test-time adaptation will be employed to mitigate domain shift across sites and cameras. To reduce labeling costs, the study will combine self-supervised pretraining, synthetic data covering extreme and long-tail conditions (e.g., BIM-rendered scenes), and active learning targeted at hard-example sampling. Finally, to ensure interpretability is preserved, future research will evaluate explanation faithfulness and stability. This includes moving beyond Grad-CAM to incorporate counterfactual and concept-based explanations, as well as developing operator interfaces that support rapid correction and feedback within the AI monitoring workflow.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961-971. https://doi.org/10.1109/CVPR.2016.110

Aken, J.E.V. Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules, Journal of Management Studies 41 (2) (2004) 219–246. https://doi.org/10.1111/j.1467-6486.2004.00430.x.

Aksu, K., & Demirel, H. (2024). Extracting structural elements from 3D point clouds in indoor environments via machine learning techniques. IEEE Access, 12, 94461–94476. https://doi.org/10.1109/ACCESS.2024.3423425

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks (arXiv:1711.06104). arXiv. https://doi.org/10.48550/arXiv.1711.06104

Angah, O., & Chen, A. Y. (2020). Tracking multiple construction workers through deep learning and the gradient-based method with re-matching based on multi-object tracking accuracy. Automation in Construction, 119, 103308. https://doi.org/10.1016/j.autcon.2020.103308

Atik, M. E., Duran, Z., & Seker, D. Z. (2024). Explainable artificial intelligence for machine learning-based photogrammetric point cloud classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 5834–5846. https://doi.org/10.1109/JSTARS.2024.3370159

Atkinson, G. A., Zhang, W., Hansen, M. F., Holloway, M. L., & Napier, A. A. (2020). Image segmentation of underfloor scenes using a Mask-RCNN with two-stage transfer learning. Automation in Construction, 113, 103118. https://doi.org/10.1016/j.autcon.2020.103118

Ayhan, B. U., & Tokdemir, O. B. (2020). Accident analysis for construction safety using latent class clustering and artificial neural networks. Journal of Construction Engineering and Management, 146, 04019114. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001762

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

- Bernardo, G. M. S., Rodrigues, J. A., & Loja, M. A. R. (2016). Towards an expeditious as-is surface reconstruction. Engineering Structures, 129, 91–107. https://doi.org/10.1016/J.ENGSTRUCT.2016.11.042

Bügler, M., Borrmann, A., Ogunmakin, G., Vela, P. A., & Teizer, J. (2017). Fusion of photogrammetry and video analysis for productivity assessment of earthwork processes. Computer-Aided Civil and Infrastructure Engineering, 32(2), 107–123. https://doi.org/10.1111/mice.12235

Cai, J., & Cai, H. (2020). Robust hybrid approach of vision-based tracking and radio-based identification and localization for 3D tracking of multiple construction workers. Journal of Computing in Civil Engineering, 34(4), 04020021. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000901

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

Chakraborty, S., Roy, M., & Hore, S. (2017). A study on different edge detection techniques in digital image processing. In Feature detectors and motion detection in video processing (pp. 100–122). IGI Global Scientific Publishing.

Chian, E., Fang, W., Goh, Y. M., & Tian, J. (2021). Computer vision approaches for detecting missing barricades. *Automation in Construction*, 131, 103862. https://doi.org/10.1016/j.autcon.2021.103862

Chan, K. C., Wang, X., Yu, K., Dong, C., & Loy, C. C. (2021). BasicVSR: The search for essential components in video super-resolution and beyond. In Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4947–4956). https://doi.org/10.48550/arXiv.2012.02181

Chattopadhay, A., Sarkar, J., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 839–847). https://doi.org/10.1109/WACV.2018.00097

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 801–818). https://doi.org/10.48550/arXiv.1802.02611

Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L. E., Ramamoorthi, R., & Su, H. (2020). Deep stereo using adaptive thin volume representation with uncertainty awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2524–2534). https://doi.org/10.48550/arXiv.1911.12012

Chi, S., Caldas, C. H., & Kim, D. Y. (2009). A methodology for object identification and tracking in construction based on spatial modeling and image matching techniques. Computer-Aided Civil and Infrastructure Engineering, 24(3), 199–211. https://doi.org/10.1111/j.1467-8667.2008.00580.x

Choi, J., Gu, B., Chin, S., & Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, 102974. https://doi.org/10.1016/j.autcon.2019.102974

Chu, M., Matthews, J., and Love, P.E.D., Integrating Mobile Building Information Modelling and Augmented Reality Systems: An Experimental Study, Automation in Construction.85 (2018) 305-316. https://doi.org/10.1016/j.autcon.2017.10.032.

Ci, H., Wang, C., & Wang, Y. (2018). Video object segmentation by learning location-sensitive embeddings. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 501–516).

Davoudi Kakhki, F., Freeman, S. A., & Mosher, G. A. (2019). Use of neural networks to identify safety prevention priorities in agro-manufacturing operations within commercial grain elevators. Applied Sciences, 9(21), 4690. https://doi.org/10.3390/app9214690

Ding, L., Fang, W., Luo, H., Love, P. E., Zhong, B., & Ouyang, X. (2018). A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. Automation in construction, 86, 118-124. https://doi.org/10.1016/j.autcon.2017.11.002

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Elhishi, S., Elashry, A. M., & El-Metwally, S. (2023). Unboxing machine learning models for concrete strength prediction using XAI. Scientific Reports, 13(1), 19892. https://doi.org/10.1038/s41598-023-47169-7

Fang, W., Ding, L., Love, P. E. D., Luo, H., Li, H., Peña-Mora, F., Zhong, B., & Zhou, C. (2020). Computer vision applications in construction safety assurance.

Automation in Construction, 110, 103013. https://doi.org/10.1016/j.autcon.2019.103013

Fang, W., Ding, L., Zhong, B., Love, P. E., & Luo, H. (2018). Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. Advanced Engineering Informatics, 37, 139–149. https://doi.org/10.1016/j.aei.2018.05.003

Fang, W., Love, P. E. D., Luo, H., & Ding, L. (2020). Computer vision for behaviour-based safety in construction: A review and future directions. Advanced Engineering Informatics, 43, 100980. https://doi.org/10.1016/j.aei.2019.100980

Fang, W., Zhong, B., Zhao, N., Love, P. E., Luo, H., Xue, J., & Xu, S. (2019). A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network. Advanced Engineering Informatics, 39, 170–177. https://doi.org/10.1016/j.aei.2018.12.005

Fogarty, G. J., & Shaw, A. (2010). Safety climate and the theory of planned behavior: Towards the prediction of unsafe behavior. Accident Analysis & Prevention, 42(5), 1455–1459. https://doi.org/10.1016/j.aap.2009.08.008

Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 3449–3457). https://doi.org/10.1109/ICCV.2017.371

Forest, F., Porta, H., Tuia, D., & Fink, O. (2024). From classification to segmentation with explainable AI: A study on crack detection and growth monitoring.

Automation in Construction, 165, 105497. https://doi.org/10.1016/j.autcon.2024.105497

Franklin, W. R. (2006). PNpoly-point inclusion in polygon test. Available at: http://www. ecse. rpi. edu/Homepages/wrf/Research/Short_Notes/pnpoly. Html, Accessed 15th December 2020

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. http://www.jstor.org/stable/2699986?origin=JSTOR-pdf

Fu, H., Meng, D., Li, W., & Wang, Y. (2021). Bridge crack semantic segmentation based on improved DeepLabv3+. Journal of Marine Science and Engineering, 9(6), 671. https://doi.org/10.3390/jmse9060671

Geerts, G.L. A design science research methodology and its application to accounting information systems research, International Journal of Accounting Information Systems. 12 (2) (2011) 142-151. https://doi.org/10.1016/j.accinf.2011.02.004.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (pp. 580–587). https://doi.org/10.1109/CVPR.2014.81

Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint. https://doi.org/10.48550/arXiv.1308.0850

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2495–2504). https://doi.org/10.48550/arXiv.1912.06378

Guo, B. H., Yiu, T. W., & González, V. A. (2016). Predicting safety behavior in the construction industry: Development and test of an integrative model. Safety Science, 84, 1–11. https://doi.org/10.1016/j.ssci.2015.11.020

Guo, W., Liang, G., Ren, S., & Zeng, C. (2024). Automatic crack detection method for high-speed railway box girder based on deep learning techniques and inspection robot. Structures, 68, 107116. https://doi.org/10.1016/j.istruc.2024.107116

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social gan: Socially acceptable trajectories with generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2255-2264). https://doi.org/10.48550/arXiv.1803.10892

Hacıefendioğlu, K., Ayas, S., Başağa, H. B., Toğan, V., Mostofi, F., & Can, A. (2022). Wood construction damage detection and localization using deep convolutional neural network with transfer learning. European Journal of Wood and Wood Products, 80(4), 791–804. https://doi.org/10.1007/s00107-022-01815-5

Health and Safety Executive. (2024). Kinds of accident statistics in Great Britain, 2024 [Report]. https://www.hse.gov.uk/statistics/assets/docs/kinds-of-accident.pdf

He, A., Luo, C., Tian, X., & Zeng, W. (2018). A twofold siamese network for real-time object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4834–4843). https://doi.org/10.48550/arXiv.1802.08817

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

Hong Kong Labour Department. (2023). Occupational safety and health statistics 2023 [Report]. https://www.labour.gov.hk/common/osh/pdf/OSH_Statistics_2023_en.pdf

International Labour Organization. (2022, September 4). ILOSTAT database: SDG labour market indicators (ILOSDG): Fatal occupational injuries per 100,000 workers [Data set]. https://ilostat.ilo.org/data/

Jia, X., Lu, H., & Yang, M. H. (2012). Visual tracking via adaptive structural local sparse appearance model. In 2012 IEEE Conference on computer vision and pattern recognition, pp. 1822-1829. https://doi.org/10.1109/CVPR.2012.6247880

Johansen, K. W., Teizer, J., & Schultz, C. (2024). Automated rule-based safety inspection and compliance checking of temporary guardrail systems in construction. Automation in Construction, 168, 105849. https://doi.org/10.1016/j.autcon.2024.105849

Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics [Data set]. GitHub. https://github.com/ultralytics/ultralytics

Kang, L. G. (2022). Statistical analysis and case investigation of fatal fall-from-height accidents in the Chinese construction industry. International Journal of Industrial Engineering: Theory, Applications and Practice, 29(3), 413–431. https://doi.org/10.23055/ijietap.2022.29.3.7971

Khan, N., Saleem, M. R., Lee, D., Park, M.-W., & Park, C. (2021). Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks. Computers in Industry, 129, 103448. https://doi.org/10.1016/j.compind.2021.103448

Kim, B., & Cho, S. (2020). Automated multiple concrete damage detection using instance segmentation deep learning model. Applied Sciences, 10(22), 8008. https://doi.org/10.3390/app10228008

Kim, D., Liu, M., Lee, S., & Kamat, V. R. (2019, May). Trajectory prediction of mobile construction resources toward pro-active struck-by hazard detection. In Proceedings of the International Symposium on Automation and Robotics in Construction (IAARC). https://doi.org/10.22260/ISARC2019/0131

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning (pp. 2668–2677). https://doi.org/10.48550/arXiv.1711.11279

Kim, D., Liu, M., Lee, S., & Kamat, V. R. (2019). Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. Automation in Construction, 99, 168–182. https://doi.org/10.1016/j.autcon.2018.12.014

Kim, H., Kim, K., & Kim, H. (2016). Vision-based object-centric safety assessment using fuzzy inference: Monitoring struck-by accidents with moving objects. Journal of Computing in Civil Engineering, 30(4), 04015075. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000562

Kim, I., Lee, Y., & Choi, J. (2020). BIM-based hazard recognition and evaluation methodology for automating construction site risk assessment. Applied Sciences, 10(7), 2335. https://doi.org/10.3390/app10072335

Kim, J., & Chi, S. (2019). Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. Automation in Construction, 104, 255–264. https://doi.org/10.1016/j.autcon.2019.03.025

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning (pp. 1885–1894). https://doi.org/10.48550/arXiv.1703.04730

Kolappan Geetha, G., & Sim, S.-H. (2022). Fast identification of concrete cracks using 1D deep learning and explainable artificial intelligence-based analysis. Automation in Construction, 143, 104572. https://doi.org/10.1016/j.autcon.2022.104572

Kolar, Z., Chen, H., & Luo, X. (2018). Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. Automation in Construction, 89, 58–70. https://doi.org/10.1016/j.autcon.2018.01.003

Kong, T., Fang, W., Love, P. E., Luo, H., Xu, S., & Li, H. (2021). Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction. Advanced Engineering Informatics, 50, 101400. https://doi.org/10.1016/j.aei.2021.101400

Kulinan, A. S., Park, M., Aung, P. P. W., Cha, G., & Park, S. (2024). Advancing construction site workforce safety monitoring through BIM and computer vision integration. Automation in Construction, 158, 105227. https://doi.org/10.1016/j.autcon.2023.105227

Kumi, L., Jeong, J., & Jeong, J. (2024). Data-driven automatic classification model for construction accident cases using natural language processing with hyperparameter tuning. Automation in Construction, 164, 105458. https://doi.org/10.1016/j.autcon.2024.105458

Lafhaj, Z., Rebai, S., AlBalkhy, W., Hamdi, O., Mossman, A., & Alves da Costa, A. (2024). Complexity in construction projects: A literature review. Buildings, 14(3), 680. https://doi.org/10.3390/buildings14030680

Li, F., Zeng, J., Huang, J., Zhang, J., Chen, Y., Yan, H., Huang, W., Lu, X., & Yip, P. S. (2019). Work-related and non-work-related accident fatal falls in Shanghai and Wuhan, China. Safety Science, 117, 43–48. https://doi.org/10.1016/j.ssci.2019.04.001

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 936–944). https://doi.org/10.1109/CVPR.2017.106

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

Liu, H., Ruan, Z., Zhao, P., Dong, C., Shang, F., Liu, Y., Yang, L., & Timofte, R. (2022). Video super-resolution based on deep learning: A comprehensive survey. Artificial Intelligence Review, 55(8), 5981–6035. https://doi.org/10.48550/arXiv.2007.12928

Liu, H., Zhao, P., Ruan, Z., Shang, F., & Liu, Y. (2021). Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), 2127–2135. https://doi.org/10.48550/arXiv.2103.11744

Liu, M., Tang, P., Liao, P.-C., & Xu, L. (2020). Propagation mechanics from workplace hazards to human errors with dissipative structure theory. Safety Science, 126, 104661. https://doi.org/10.1016/j.ssci.2020.104661

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European Conference on Computer Vision (pp. 21–37). https://doi.org/10.48550/arXiv.1512.02325

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 150–158). https://doi.org/10.1145/2339530.2339556

Love, P. E. D., Fang, W., Matthews, J., Porter, S., Luo, H., & Ding, L. (2023). Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. Advanced Engineering Informatics, 57, 102024. https://doi.org/10.1016/j.aei.2023.102024

Love, P. E., Teo, P., & Morrison, J. (2018). Unearthing the nature and interplay of quality and safety in construction projects: An empirical study. Safety Science, 103, 270–279. https://doi.org/10.1016/j.ssci.2017.11.026

Love, P. E., Teo, P., Ackermann, F., Smith, J., Alexander, J., Palaneeswaran, E., & Morrison, J. (2018). Reduce rework, improve safety: An empirical inquiry into the precursors to error in construction. Production Planning & Control, 29(5), 353–366. https://doi.org/10.1080/09537287.2018.1424961

Lu, W., Qian, M., Xia, Y., Lu, Y., Shen, J., Fu, Q., & Lu, Y. (2024). Crack _ PSTU: Crack detection based on the U-Net framework combined with Swin Transformer. Structures, 62, 106241. https://doi.org/10.1016/j.istruc.2024.106241

Lu, Y., Gong, P., Tang, Y., Sun, S., & Li, Q. (2021). BIM-integrated construction safety risk assessment at the design stage of building projects. Automation in Construction, 124, 103553. https://doi.org/10.1016/j.autcon.2021.103553

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30.

Luo, H., Chen, J., Love, P. E. D., & Fang, W. (2024). Explainable transfer learning for modeling and assessing risks in tunnel construction. IEEE Transactions on Engineering Management, 71, 8339–8355. https://doi.org/10.1109/TEM.2024.3369231

Luo, H., Liu, J., Fang, W., Love, P. E., Yu, Q., & Lu, Z. (2020). Real-time smart video surveillance to manage safety: A case study of a transport mega-project. Advanced Engineering Informatics, 45, 101100.

Luo, J.-Y., & Liu, Y.-C. (2024). Adaptive and explainable deep learning-based rapid identification of architectural cracks. IEEE Access, 12, 111741–111751. https://doi.org/10.1109/ACCESS.2024.3442926

Luo, X., Li, H., Cao, D., Dai, F., Seo, J., & Lee, S. (2018). Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. Journal of Computing in Civil Engineering, 32(3), 04018012. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000756

Luo, X., Li, H., Wang, H., Wu, Z., Dai, F., & Cao, D. (2019). Vision-based detection and visualization of dynamic workspaces. Automation in Construction, 104, 1–13. https://doi.org/10.1016/j.autcon.2019.04.001

Ma, C., Guo, Q., Jiang, Y., Luo, P., Yuan, Z., & Qi, X. (2022). Rethinking resolution in the context of efficient video recognition. In Advances in Neural Information

Processing Systems, 35, 37865–37877. https://doi.org/10.48550/arXiv.2209.12797

Macrae, C. (2022). Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. Risk Analysis, 42(9), 1999–2025. https://doi.org/10.1111/risa.13850

Man, S. S., Chan, A. H. S., Alabdulkarim, S., & Zhang, T. (2021). The effect of personal and organizational factors on the risk-taking behavior of Hong Kong construction workers. Safety Science, 136, 105155. https://doi.org/10.1016/j.ssci.2020.105155

Martinez, J. G., Gheisari, M., & Alarcon, L. F. (2020). UAV integration in current construction safety planning and monitoring processes: Case study of a high-rise building construction project in Chile. Journal of Management in Engineering, 36(3), 05020005. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000761

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Ministry of Housing and Urban–Rural Development (MOHURD). (2022). Document for supervision punishment. https://www.mohurd.gov.cn/gongkai/zhengce/zhengcefilelib/202210/20221026_768565.html

Mirzaei, K., Arashpour, M., Asadi, E., Masoumi, H., Bai, Y., & Behnood, A. (2022). 3D point cloud data processing with machine learning for construction and infrastructure applications: A comprehensive review. Advanced Engineering Informatics, 51, 101501. https://doi.org/10.1016/j.aei.2021.101501

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Na, X. U., Ling, M. A., & Li, W. (2021). An improved text mining approach to extract safety risk factors from construction accident reports. Safety Science, 138, 105216. https://doi.org/10.1016/j.ssci.2021.105216

Neuhausen, M., Teizer, J., & König, M. (2018). Construction worker detection and tracking in bird's-eye view camera images. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, 35, 1–8. https://doi.org/10.22260/ISARC2018/0161

Newaz, M. T., Ershadi, M., Carothers, L., Jefferies, M., & Davis, P. (2022). A review and assessment of technologies for addressing the risk of falling from height on construction sites. Safety Science, 147, 105618. https://doi.org/10.1016/j.ssci.2021.105618

Oliveira, S. S., De Albuquerque Soares, W., & Vasconcelos, B. M. (2023). Fatal fall-from-height accidents: Statistical treatment using the Human Factors Analysis and Classification System–HFACS. Journal of Safety Research, 86, 118–126. https://doi.org/10.1016/j.jsr.2023.05.004

Patel, D. A., & Jha, K. N. (2015). Neural network model for the prediction of safe work behavior in construction projects. *Journal of Construction Engineering and Management*, 141(1), 04014066. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000922

Park, C.-S., & Kim, H.-J. (2013). A framework for construction safety management and visualization system. Automation in Construction, 33, 95–103. https://doi.org/10.1016/j.autcon.2012.09.012

Peerally, M. F., Carr, S., Waring, J., & Dixon-Woods, M. (2017). The problem with root cause analysis. BMJ Quality & Safety, 26, 417. https://doi.org/10.1136/bmjqs-2016-005511

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 724-732). https://doi.org/10.1109/CVPR.2016.85

Perazzi, F., Khoreva, A., Benenson, B., Schiele, B., & Sorkine-Hornung, A. (2017). Learning video object segmentation from static images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2663–2672). https://doi.org/10.1109/CVPR.2017.372

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models (arXiv:1806.07421). arXiv. https://doi.org/10.48550/arXiv.1806.07421

Pham, K.-T., Vu, D.-N., Hong, P. L. H., & Park, C. (2020). 4D-BIM-based workspace planning for temporary safety facilities in construction SMEs. International Journal of Environmental Research and Public Health, 17(10), 3403. https://doi.org/10.3390/ijerph17103403

Pinheiro, P. O., Collobert, R., and Dollár, P. (2015). *Learning to segment object candidates*. In Advances in Neural Information Processing Systems pp. 1990-1998. https://doi.org/10.48550/arXiv.1506.06204

Pinheiro, P. O., Lin, T. Y., Collobert, R., and Dollár, P. (2016). *Learning to refine object segments*. In European Conference on Computer Vision, Springer, Cham, pp.75-91. https://doi.org/10.48550/arXiv.1603.08695

Pruthi, G., Liu, F., Kale, S., & Sundararajan, M. (2020). Estimating training data influence by tracing gradient descent. In Advances in Neural Information Processing Systems, 33, 19920–19930. https://doi.org/10.48550/arXiv.2002.08484

Qi, H., Zhou, Z., Li, N., & Zhang, C. (2022). Construction safety performance evaluation based on data envelopment analysis (DEA) from a hybrid perspective of cross-sectional and longitudinal. Safety Science, 146, 105532. https://doi.org/10.1016/j.ssci.2021.105532

Qi, H., Zhou, Z., Manu, P., & Li, N. (2025). Falling risk analysis at workplaces through an accident data-driven approach based upon hybrid artificial intelligence (AI) techniques. Safety Science, 185, 106814. https://doi.org/10.1016/j.ssci.2025.106814

Qi, J., Issa, R. R., Olbina, S., & Hinze, J. (2014). Use of building information modeling in design to prevent construction worker falls. Journal of Computing in Civil Engineering, 28(5), A4014008. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000365

Rafindadi, A. D., Shafiq, N., & Othman, I. (2022). A conceptual framework for BIM process flow to mitigate the causes of fall-related accidents at the design stage. Sustainability, 14(20), 13025. https://doi.org/10.3390/su142013025

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, 28. https://doi.org/10.1109/TPAMI.2016.2577031.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Riedel, H., Mokdad, S., Schulz, I., Kocer, C., Rosendahl, P. L., Schneider, J., Kraus, M. A., & Drass, M. (2022). Automated quality control of vacuum insulated glazing by convolutional neural network image classification. Automation in Construction, 135, 104144. https://doi.org/10.1016/j.autcon.2022.104144

Roberts, D., & Golparvar-Fard, M. (2019). End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. Automation in Construction, 105, 102811. https://doi.org/10.1016/j.autcon.2019.04.006

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2006). Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering & System Safety, 91(10), 1109–1125. https://doi.org/10.1016/j.ress.2005.11.014

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). https://10.1109/ICCV.2017.74

Seo, J., Yin, K., and Lee, S. (2016). *Automated postural ergonomic assessment using a computer vision-based posture classification*. In Construction Research Congress, pp. 809-818. https://doi.org/10.1016/j.autcon.2021.103725

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. Journal of Consumer Research, 325–343. http://dx.doi.org/10.1086/209170

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning (pp. 3145–3153).

Stemn, E., Bofinger, C., Cliff, D., & Hassall, M. E. (2018). Failure to learn from safety incidents: Status, challenges and opportunities. Safety science, 101, 313-325. https://doi.org/10.1016/j.ssci.2017.09.018

Spielberger, C. (Ed.). (2004). Encyclopedia of applied psychology. Academic Press.

Sun, H., Lu, D., Li, X., Tan, J., Zhao, J., & Hou, D. (2024). Research on multi-apparent defects detection of concrete bridges based on YOLOR. Structures, 65, 106735. https://doi.org/10.1016/j.istruc.2024.106735

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 3319–3328).

Sulowski, A. (2014). Collective fall protection for construction workers. Informes de la Construcción, 66(533). https://doi.org/10.3989/ic.12.035

Tan, Y., Li, G., Cai, R., Ma, J., & Wang, M. (2022). Mapping and modelling defect data from UAV captured images to BIM for building external wall inspection. *Automation in Construction*, 139, 104284. https://doi.org/10.1016/j.autcon.2022.104284

Tariq, A., Ali, B., Ullah, F., & Alqahtani, F. K. (2023). Reducing falls from heights through BIM: A dedicated system for visualizing safety standards. Buildings, 13(3), 671. https://doi.org/10.3390/buildings13030671

U.S. Bureau of Labor Statistics. (2022). Occupational injuries/illness and fatal injuries profiles. https://data.bls.gov/gqt/InitialPage

Uddin, S. M., Albert, A., Alsharef, A., Pandit, B., Patil, Y., & Nnaji, C. (2020). Hazard recognition patterns demonstrated by construction workers. International Journal of Environmental Research and Public Health, 17(21), 7788. https://doi.org/10.3390/ijerph17217788

Wang, F., Galliani, S., Vogel, C., Speciale, P., & Pollefeys, M. (2021). PatchMatchNet: Learned multi-view PatchMatch stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14194–14203). https://doi.org/10.1109/CVPR46437.2021.01397

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 24–25).

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., and Torr, P. H. (2019). *Fast online object tracking and segmentation: A unifying approach*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 1328-1338. https://doi.org/10.48550/arXiv.1812.05050

Wang, J., Zhang, S., & Teizer, J. (2015). Geotechnical and safety protective equipment planning using range point cloud data and rule checking in building information modeling. Automation in Construction, 49, 250–261. https://doi.org/10.1016/j.autcon.2014.09.002

Wang, Q. (2019). Automatic checks from 3D point cloud data for safety regulation compliance for scaffold work platforms. Automation in Construction, 104, 38–51. https://doi.org/10.1016/j.autcon.2019.04.008

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. (2019). Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1328–1338). https://doi.org/10.48550/arXiv.1812.05050

Wang, T., & Gan, V. J. L. (2024). Multi-view stereo for weakly textured indoor 3D reconstruction. Computer-Aided Civil and Infrastructure Engineering, 39(10), 1469–1489. https://doi.org/10.1111/mice.13149

Wang, W., Su, C., & Fu, D. (2022). Automatic detection of defects in concrete structures based on deep learning. Structures, 43, 192–199. https://doi.org/10.1016/j.istruc.2022.06.042

Wei, F., Yao, G., Yang, Y., & Sun, Y. (2019). Instance-level recognition and quantification for concrete surface bughole based on deep learning. Automation in Construction, 107, 102920. https://doi.org/10.1016/j.autcon.2019.102920

Wu, H., Zhong, B., Li, H., Love, P., Pan, X., & Zhao, N. (2021). Combining computer vision with semantic reasoning for on-site safety management in construction. Journal of Building Engineering, 42, 103036. https://doi.org/10.1016/j.jobe.2021.103036

Wu, P., Liu, A., Fu, J., Ye, X., & Zhao, Y. (2022). Autonomous surface crack identification of concrete structures based on an improved one-stage object

detection algorithm. Engineering Structures, 272, 114962. https://doi.org/10.1016/j.engstruct.2022.114962

Xiao, B., & Zhu, Z. (2018). Two-dimensional visual tracking in construction scenarios: A comparative study. *Journal of Computing in Civil Engineering,* 32(3), 04018006.

An, X., Zhou, L., Liu, Z., Wang, C., Li, P., & Li, Z. (2021). Dataset and benchmark for detecting moving objects in construction sites. Automation in Construction, 122, 103482. https://doi.org/10.1016/j.autcon.2020.103482

Yan, Y., Zhang, S., Wang, X., & Li, X. (2024). Novel unsupervised machine learning method for identifying falling from height hazards in building information models through path simulation sampling. Advances in Civil Engineering, 2024(1), 6333621. https://doi.org/10.1155/2024/6333621

Yang, X., Li, H., Yu, Y., Luo, X., Huang, T., & Yang, X. (2018). Automatic pixel-level crack detection and measurement using fully convolutional network. Computer-Aided Civil and Infrastructure Engineering, 33(12), 1090–1109. https://doi.org/10.1111/mice.12412

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., & Quan, L. (2019). Recurrent MVSNet for high-resolution multi-view stereo depth inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5525–5534). https://doi.org/10.1109/CVPR.2019.00567

Ye, G., Qu, J., Tao, J., Dai, W., Mao, Y., & Jin, Q. (2023). Autonomous surface crack identification of concrete structures based on the YOLOv7 algorithm. Journal of Building Engineering, 73, 106688. https://doi.org/10.1016/j.jobe.2023.106688

Yong, G., Liu, M., & Lee, S. (2024). Explainable image captioning to identify ergonomic problems and solutions for construction workers. Journal of Computing in Civil Engineering, 38(4), 04024022. https://doi.org/10.1061/JCCEE5.CPENG-5744

Yuan, W., Wang, J., & Xu, W. (2022). Shift pooling PSPNet: Rethinking PSPNet for building extraction in remote sensing images from entire local feature pooling. Remote Sensing, 14(19), 4889. https://doi.org/10.3390/rs14194889

Zeng, Q., Fan, G., Wang, D., Tao, W., & Liu, A. (2024). A systematic approach to pixel-level crack detection and localization with a feature fusion attention network and 3D reconstruction. Engineering Structures, 300, 117219. https://doi.org/10.1016/j.engstruct.2023.117219

Zhang, S., Sulankivi, K., Kiviniemi, M., Romo, I., Eastman, C. M., & Teizer, J. (2015). BIM-based fall hazard identification and prevention in construction safety planning. Safety Science, 72, 31–45. https://doi.org/10.1016/j.ssci.2014.08.001

Zhang, S., Teizer, J., Lee, J.-K., Eastman, C. M., & Venugopal, M. (2013). Building information modeling (BIM) and safety: Automatic safety checking of construction models and schedules. Automation in Construction, 29, 183–195. https://doi.org/10.1016/j.autcon.2012.05.006

Zhang, W., Zhu, S., Zhang, X., & Zhao, T. (2020). Identification of critical causes of construction accidents in China using a model based on system thinking and case analysis. Safety Science, 121, 606–618. https://doi.org/10.1016/j.ssci.2019.04.038

Zhao, Y., Zhang, Y., Xu, J., Yu, P., Wu, J., Deng, S., & Song, X. (2020). Seismic performance of the frame-skin structure based on 3D scanning technology. Structures, 23, 789–798. https://doi.org/10.1016/j.istruc.2019.12.015

Zhe, Y., Hou, K., Wang, Z., Yang, S., & Sun, H. (2025). Prediction model for grouting volume using borehole image features and explainable artificial intelligence. Construction and Building Materials, 470, 140626. https://doi.org/10.1016/j.conbuildmat.2025.140626

Zhong, W., Lu, H., & Yang, M. H. (2014). Robust object tracking via sparse collaborative appearance model. IEEE Transactions on Image Processing, 23(5), 2356-2368. https://doi.org/10.1109/TIP.2014.2313227

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2921–2929). https://doi.org/10.48550/arXiv.1512.04150

Zhou, Z., Yu, X., Zhu, Z., Zhou, D., & Qi, H. (2023). Development and application of a Bayesian network-based model for systematically reducing safety risks in the commercial air transportation system. Safety Science, 157, 105942. https://doi.org/10.1016/j.ssci.2022.105942

Zhu, Z., Ren, X., & Chen, Z. (2016). Visual tracking of construction jobsite workforce and equipment with particle filtering. Journal of Computing in Civil Engineering, 30(6), 04016023. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000573

Zhu, Z., Ren, X., & Chen, Z. (2017). Integrated detection and tracking of workforce and equipment from construction jobsite videos. Automation in Construction, 81, 161–171. https://doi.org/10.1016/j.autcon.2017.05.005