DEEP LEARNING-BASED POSE ESTIMATION FOR IDENTIFYING POTENTIAL FALL HAZARDS OF CONSTRUCTION WORKER

Minsoo Park

Sungkyun AI Research Institute, Sungkyunkwan University, South Korea

Seungsoo Lee, Woonggyu Choi & Yuntae Jeon

Department of Global Smart City, Sungkyunkwan University, South Korea

Dai Quoc Tran

Global Engineering Institute for Ultimate Society, Sungkyunkwan University, South Korea

Seunghee Park

School of Civil, Architectural Engineering and Landscape Architecture, Sungkyunkwan University, South Korea

ABSTRACT: *Fall from height (FFH) is one of the major causes of injury and fatalities in construction industry. Deep learning-based computer vision for safety monitoring has gained attention due to its relatively lower initial cost compared to traditional sensing technologies. However, a single detection model that has been used in many related studies cannot consider various contexts at the construction site. In this paper, we propose a deep learning-based pose estimation approach for identifying potential fall hazards of construction workers. This approach can relatively increase the accuracy of estimating the distance between the worker and the fall hazard area compared to the existing methods from the experimental results. Our proposed approach can improve the robustness of worker location estimation compared to existing methods in complex construction site environments with obstacles that can obstruct the worker's position. Also, it is possible to provide information on whether a worker is aware of a potential fall risk area. Our approach can contribute to preventing FFH by providing access information to fall risk areas such as construction site openings and inducing workers to recognize the risk area even in Inattentional blindness (IB) situations.*

KEYWORDS: *deep learning, keypoint detection, pose estimation, computer vision, construction site safe*

1. INTRODUCTION

Due to numerous hazards and safety challenges, the construction industry stands out as highly dangerous, characterized by elevated rates of accidents and injuries. Among these, falls from heights (FFH) emerge as a particularly frequent and urgent concern, often leading to severe injuries or fatal outcomes. These incidents underscore the inherent risks associated with construction activities, contributing to delays and economic setbacks (Rafindadi et al., 2022).

Despite the stringent enforcement of safety standards, comprehensive worker training, and the adoption of advanced protective equipment, FFH-related accidents persist at an alarming rate. A closer examination reveals that these mishaps frequently result from worker negligence, inadequate situational awareness, or an inability to recognize impending dangers (Golparvar-Fard et al., 2013). The dynamic and ever-evolving nature of construction sites further exacerbates these challenges, rendering many traditional safety measures ineffective.

Historically, human oversight and routine inspections have been the primary means of safety supervision in construction settings. However, these methods, being inherently subjective, often result in inconsistent safety assessments. Recognizing these limitations, there's been a shift towards leveraging emerging technologies such as computer vision and artificial intelligence (AI). While these innovations promise objective, consistent, and real-time safety evaluations, challenges remain. Specifically, detecting hazards like floor openings becomes complex due to occlusions from construction materials and scaffolding. Additionally, determining a worker's position, especially when parts of their body are obscured, remains problematic.

In light of these challenges, this study proposes a novel approach, integrating computer vision and deep learning, tailored for construction site safety evaluations. The essence of our methodology lies in the fusion of quadrilateral detection, pose estimation, and single depth estimation. Quadrilateral detection accurately captures the contours

Referee List (DOI: 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Minsoo Park, Seungsoo Lee, Woonggyu Choi, Yuntae Jeon, Dai Quoc Tran, Seunghee Park, *Deep Learning-Based Pose Estimation for Identifying Potential Fall Hazards of Construction Worker,* pp. 634-640, © 2023 Author(s), CC BY NC 4.0, DOI 10.36253/979-12-215-0289-3.62

of target objects, pose estimation provides insights into their spatial orientation, and single depth estimation refines the distance measurements. By employing quadrilateral anchors, further enhanced by the Vision Transformer, our approach aims to provide a more robust and accurate tool for hazard detection and prevention.

The structure of this paper is as follows:

- Chapter 2 delves into relevant literature, providing a thorough assessment of existing approaches, their strengths, and inherent limits.
- Chapter 3 elucidates our proposed methodology, shedding light on its unique facets and potential advantages.
- Chapter 4 includes experimental results, including quantitative assessments as well as visual representations of our findings.
- Chatpter 5 concludes the paper by summarizing the paper and suggesting future avenues for research.

2. RELATED WORK

The construction industry has long grappled with the challenge of ensuring worker safety, especially in the context of falls from heights (FFH) (Helander, M. 1980). This section delves into the existing methods and recent advancements in recognizing unsafe areas and the application of deep learning in the construction site context.

2.1. FFH-related Safety Monitoring

Traditional safety measures, such as guardrails and safety nets, have been the primary defense against FFH incidents (Zhang, M. and Fang, D. 2013). However, the dynamic and complex nature of construction sites often renders these measures insufficient. The rapidly changing environment, coupled with the diverse nature of construction tasks, necessitates more advanced and adaptive safety solutions.

Historically, the realm of automated construction safety has leaned heavily on sensor-based mechanisms. Techniques like radio frequency identification (RFID) were the go-to solutions for monitoring workers' movements into potentially hazardous zones (Costin, et al., 2012). Similarly, tools like global positioning systems (GPS) and ultra-wideband (UWB) played pivotal roles in identifying unsafe regions and pinpointing the location of workers and materials (Pradhananga, N. and Teizer, J. 2013). However, the drawback of these methods was the necessity for individual sensor installations.

The modern era has seen a surge in the exploration of computer vision combined with deep learning as potential reasonable approach in safety area (Park, et al., 2020; Jeon, et al., 2023). Recent advancements in technology are poised to significantly transform the paradigm of worker safety through the automation of risk assessments. Noteworthy progress has been achieved in the application of computer vision to identify potential hazards (Tran, et al., 2022). However, some studies exhibit limitations in their scope, especially regarding the precise localization of dangers and the evaluation of workers' proximity to such hazards. In sight of these observations, there's a clear demand for a more refined approach that not only pinpoints hazards with precision but also factors in worker proximity in real-world scenarios.

2.2. Deep Learning and Computer Vision in Construction Environments

The dynamic and cluttered nature of construction sites poses unique challenges for deep learning models. The presence of obstructions like construction materials and scaffolding often disrupts the model's ability to accurately identify target objects. A popular solution, borrowed from interdisciplinary research, is the integration of attention modules. These modules enhance the model's feature extraction capabilities, emphasizing crucial aspects of images.

Recent research has showcased the potential of attention mechanisms in improving detection accuracy, especially in environments where objects are either partially hidden or appear smaller due to perspective. For instance, several works refined the triplet attention mechanism, assigning greater significance to vital features. This refinement allowed models to zero in on specific image sections, enhancing the accuracy of worker detection, even in intricate construction settings. Their model could pinpoint a worker's approximate location, even if they were partially obscured. Another noteworthy approach is the use of distance intersection over union (DIoU) based on non-maximum suppression (NMS). This technique distinguishes between overlapping objects against the complex backdrop of construction sites.

However, these methods are not without their limitations. For instance, if a worker's lower body is entirely obscured, the detection only captures the visible sections. This makes it challenging to determine a worker's exact position based solely on bounding box coordinates.

Enter pose estimation, a more adaptable solution to traditional object detection challenges posed by occlusions. Recent advancements in the intersection of construction safety and computer vision have highlighted its promise. By leveraging pose estimation, researchers have been able to identify unsafe work postures and ensure the proper use of safety equipment.

Despite its potential, few have explored pose estimation to determine construction worker locations, especially considering obscured body parts. Moreover, basic detection methods, which don't account for distance, only identify the presence of workers and hazards without gauging the relative proximity between them. This limitation hampers their ability to issue timely warnings to workers approaching danger zones.

The Vision Transformer model, applied to depth estimation, demonstrated the capability of assessing object depths and distances with a singular camera (Ranftl, et al., 2021). However, with increasing distances, the differentiation in depth becomes less discernible, potentially limiting its utility for comprehensive construction site surveillance. In conclusion, there is a pressing need for cost-effective computer vision techniques that can not only pinpoint worker locations but also preemptively warn them about impending hazards.

3. METHODS

To address the identified challenges, we propose a comprehensive methodology that leverages advanced computer vision and deep learning techniques. Our methodology consists of two main components: detection of floor openings using a quadrilateral-anchor-based Vision Transformer, and estimation of worker positions using a pose estimation approach.

3.1 Baseline Detection Architecture

Given the urgent and real-time nature of risk management on construction sites, it is imperative to employ a model that can provide immediate and accurate monitoring. YOLOv7 is boasting high accuracy while maintaining real-time capabilities in real time detection (Wang, et al., 2023). While minor trade-offs in FPS might occur, extending YOLOv7 with additional models presents an avenue for enhancing accuracy. In view of these considerations, YOLOv7 was chosen as the foundational model for our study.

3.2 Attention Mechanism

The introduction of the attention mechanism has proven transformative in object detection tasks. By dynamically emphasizing salient image features while downplaying less informative ones, models exhibit enhanced capacity for accurate object detection and classification, all while preserving the efficiency of the detection process (Park, et al., 2022; Guo, et al., 2022). Traditional attention methods, including squeeze-and-excitation (SE) (Hu, et al., 2018) and convolutional block attention mechanism (CBAM) (Woo, et al., 2018), utilize convolutional neural networks to recalibrate feature maps, enhancing model accuracy and robustness by prioritizing meaningful information over less relevant components.

3.3 Polygon Anchor for Object Detection with Vision Transformer

The detection of floor openings is achieved through a convex quadrilateral-anchor-based Vision Transformer. This approach allows for accurate detection of floor openings, even when the camera perspective is not aligned with the floor opening as shown in Figure 1. The quadrilateral anchors allow for more flexibility in defining the bounding boxes, enabling them to closely match the actual boundaries of the floor openings. The Vision Transformer is used to extract both the global dependencies between the floor openings and other parts of the building and the local features specific to the floor openings. This combination of global and local feature extraction enhances the detection performance of the floor openings.

These two components, when combined, allow for a comprehensive, real-time monitoring of safety on construction sites. They enable not only the detection of floor openings but also the identification of unsafe zones and the real-time tracking of worker positions. This comprehensive safety monitoring can significantly enhance the safety at construction sites.

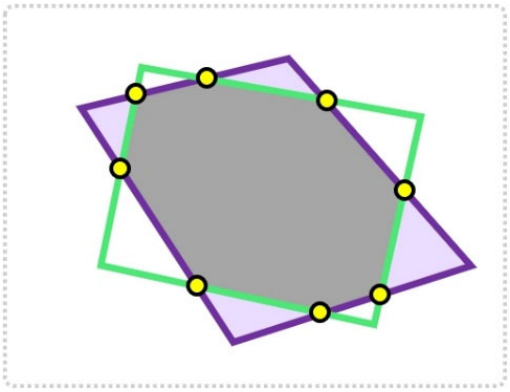

Fig. 1: Using convex quadrilateral anchors to detect floor openings. The purple boxes represent the ground truth, while the green boxes represent the predicted boxes. The overlapping area between the two boxes, indicated by yellow dots, is used to calculate the IoU (Intersection over Union).

3.4 Integrated Model for Floor-Opening Detection

we incorporate the YOLO-pose model (Maji, et al., 2022), which enhances YOLO capability to predict human poses. This model leverages pose estimation to refine object localization, offering a more comprehensive understanding of worker positions and orientations in relation to floor openings. This integration aims to amplify the focus on critical features and supply a more comprehensive contextual representation of images. This augmentation potentially heightens the efficacy and precision of our object detection model.

The envisioned integrated model aims to capture both local attributes and global interdependencies of floor openings. Local characteristics encompass attributes like shape, size, and color, facilitating nuanced comprehension and enabling their integration with other objects. Meanwhile, global interdependencies offer a holistic understanding of image constituents. For instance, floor openings often correlate with specific building elements, such as slabs, and a global outlook can encapsulate contextual information on a broader scale. This approach not only bolsters the perception of individual components but also provides insights into the overall building structure and floor opening placements.

3.5 Estimation of Relative Distances Between Workers and Danger Zone

Ensuring safety on construction sites entails assessing the proximity of workers to potential hazards, in this case, openings or drop-offs. Our method employs a combination of pose estimation and depth estimation to achieve this.

Firstly, we leverage the pose estimation of workers, specifically focusing on the leg parts, to gauge the distance, denoted as D. The rationale behind this focus is that the legs are often the closest body parts to openings or drop-offs and thus serve as critical indicators of a worker's proximity to these hazards.

Once D is determined, we then identify areas within a radius of D from detected openings as "hazard zones." Any worker located within this zone is considered at risk, warranting immediate safety interventions. Also, In the context of drop-offs, we utilize single depth estimation. When a detected individual's depth estimation result exhibits a drastic change, indicating proximity to areas with significant depth differences, they are deemed to be in a hazardous zone. Essentially, if a worker is close to an area where the depth changes abruptly, it is considered a potential fall hazard.

By integrating pose and depth estimations, our approach provides a comprehensive measure of potential risks, enabling proactive safety measures on construction sites.

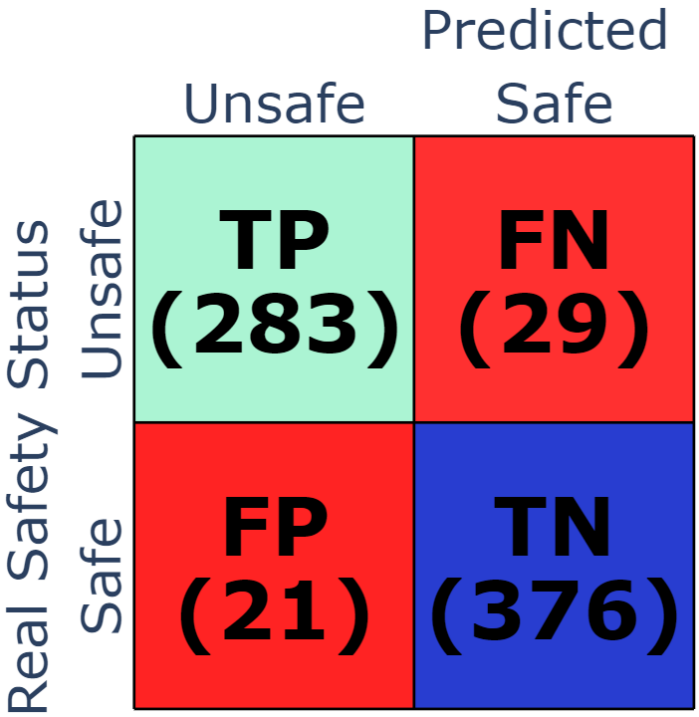

Fig. 2: The results of confusion matrix for the safe/unsafe decision making in construction site.

4. RESULTS

For our experiments, data were collected from both real construction sites and 3D simulation models. We specifically gathered images and videos of workers situated near openings and drop-offs where safety measures were not adequately implemented. In total, 3545 datasets were collected. We then employed a training-validation-test schema with a split ratio of 3:1:1, respectively, to evaluate our model's performance. The proposed method not only improves the detection of floor openings but also estimates the relative distance between the workers and the openings. The method defines unsafe zones around the openings based on this relative distance and provides real-time warnings to the workers when they enter these zones. The experiments also demonstrate the robustness of the proposed method in handling the complex and dynamic environment of construction sites.

The quantitative accuracy of the alert for FFH prevention at openings and edges from the proposed method is represented in Fig. 2, visualized using a Confusion matrix. We used a confusion matrix to assess the performance of our model in predicting whether a worker is in a hazardous zone or a safe zone. In real construction sites, due to the inherent nature of the environment, most of the samples were from safe zones, which explains the higher number of safe situation. By testing in both real and virtual environments, we ensured a comprehensive evaluation of our model's performance across diverse scenarios. Our model achieved an accuracy of approximately 93.2%, representing the proportion of total predictions that were correct. The precision of the model was 93.1%, indicating that when our model predicted a worker to be in a hazardous zone, it was correct about 93.1% of the time. The recall was valued at 90.7%, showcasing that our model correctly identified 90.7% of all actual hazardous situations. Harmonizing precision and recall, the F1-Score was found to be 91.9%. Fig. 3 visualizes the inference results of the models used in the proposed methodology. The final decision on risk/safety is automatically determined

through post-processing from a combination of three algorithms.

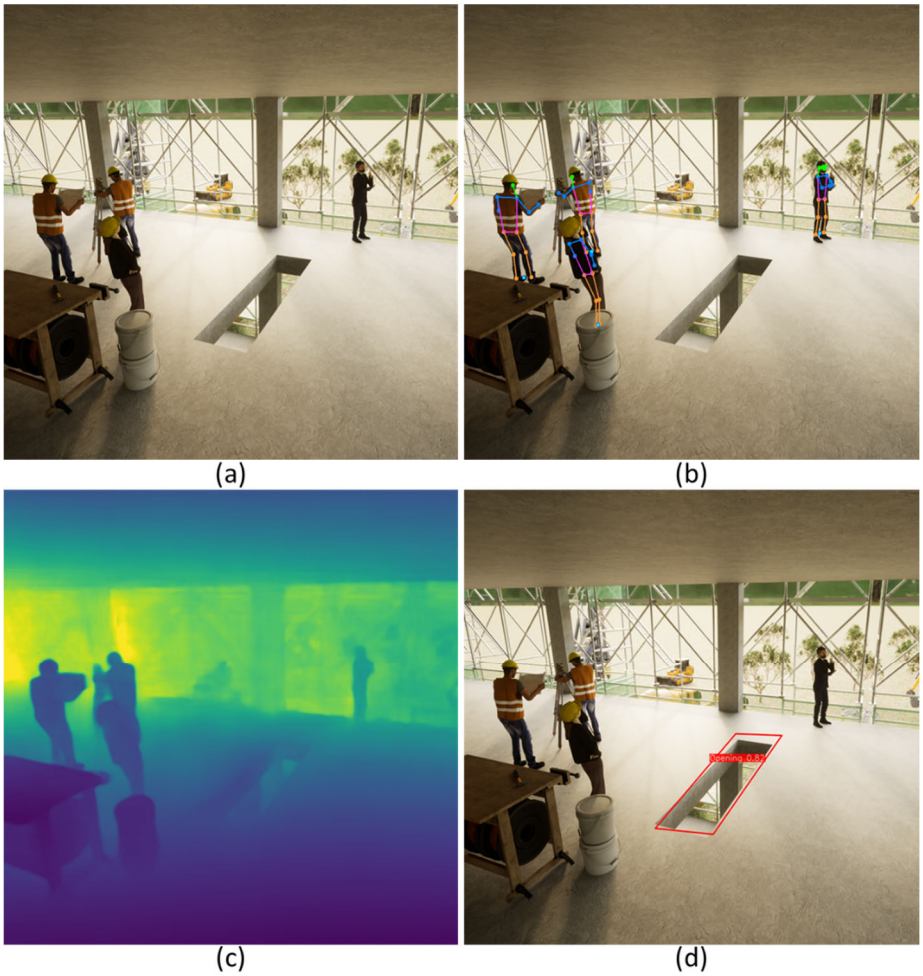

Fig. 3: Results of computer vision models in simulated hazardous situations for openings and edges using Unity. (a) Original image, (b) Pose estimation results, (c) Depth estimation results from a single camera, (d) Polygon detection results for floor openings.

5. CONCLUSIONS

This study presents a pivotal advancement in automated safety monitoring within the construction domain. By harnessing state-of-the-art computer vision and deep learning paradigms, our method adeptly detects floor openings and estimates the proximity of workers, facilitating real-time risk assessment. The potential to preempt accidents and heighten safety through this methodology is profound.

However, our study acknowledges certain limitations. Notably, while our method is primed for detecting risks, it may register false positives, especially in scenarios where floor openings function as stairs or access points, and workers are expectedly moving in and out. Conversely, false negatives can manifest when workers operate beyond a specific distance from the camera, rendering them undetectable. As per the feedback, it's crucial to clarify that we haven't explicitly stated or tested the exact distance threshold on-site, which denotes the limit beyond which workers might not be detected. These limitations point to the need for further research and improvement in our method. Future work should aim to address these issues, as well as incorporate considerations of existing safety measures on construction sites to provide more nuanced safety monitoring information.

6. ACKNOWLEDGEMNT

This research was supported by a grant [2022-MOIS38-002 (RS-2022-ND630021)] from the Ministry of Interior and Safety (MOIS)’s project for the development of accident prevention technology for vulnerable groups. In

addition, this research was supported by a grant from the Korean Government (MSIT) to the Research Foundation of Korea (NRF) [RS-2023-00250166] and This work is financially supported by Korea Ministry of Land, Infrastructure and Transport(MOLIT) as 「Innovative Talent Education Program for Smart City.

REFERENCES

Costin, A., Pradhananga, N., & Teizer, J. (2012). Leveraging passive RFID technology for construction resource field mobility and status monitoring in a high-rise renovation project. *Automation in Construction*, 24, 1-15.

Golparvar-Fard, M., Heydarian, A., & Niebles, J. C. (2013). Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, 27(4), 652-663.

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3), 331-368.

Helander, M. (1980). Safety challenges in the construction industry. *Journal of Occupational Accidents*, 2(4), 257-263.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132-7141.

Jeon, Y., Tran, D. Q., Park, M., & Park, S. (2023). Leveraging Future Trajectory Prediction for Multi-Camera People Tracking. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5398-5407.

Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2637-2646.

Park, M., Tran, D. Q., Dai. Bak, J., & Park, S. (2023). Small and overlapping worker detection at construction sites. *Automation in Construction*, 151, 104856.

Park, M., Tran, D. Q., Jung, D., & Park, S. (2020). Wildfire-detection method using DenseNet and CycleGAN data augmentation-based remote camera imagery. *Remote Sensing*, 12(22), 3715.

Park, M., Tran, D. Q., Bak, J., & Park, S. (2022). Advanced wildfire detection using generative adversarial network-based augmented datasets and weakly supervised object localization. *International Journal of Applied Earth Observation and Geoinformation*, 114, 103052.

Pradhananga, N., & Teizer, J. (2013). Automatic spatio-temporal analysis of construction site equipment operations using GPS data. *Automation in construction*, 29, 107-122.

Rafindadi, A. D. U., Napiah, M., Othman, I., Mikić, M., Haruna, A., Alarifi, H., & Al-Ashmori, Y. Y. (2022). Analysis of the causes and preventive measures of fatal fall-related accidents in the construction industry. *Ain Shams Engineering Journal*, 13(4), 101712.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. *In Proceedings of the IEEE/CVF international conference on computer vision*, 12179-12188.

Tran, D. Q., Park, M., Jeon, Y., Bak, J., & Park, S. (2022). Forest-fire response system using deep-learning-based approaches with CCTV images and weather data. *IEEE Access*, 10, 66061-66071.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464-7475.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *In Proceedings of the European conference on computer vision (ECCV)*, 3-19.

Zhang, M., & Fang, D. (2013). A cognitive analysis of why Chinese scaffolders do not use safety harnesses in construction. *Construction Management and Economics*, 31(3), 207-222.