A Real-time Intelligent Scene Understanding Systems for Human-Centered Interaction in Dynamic Industrialized Construction

Linfeng Richard Xu[a], Aladdin Alwisy[b,*]

[a]*ISE, University of Florida, Gainesville, FL*
[b]*DCP, University of Florida, Gainesville, FL*

Abstract

Construction processes demand enhanced safety, efficiency, and task clarity to accommodate the increasing complexity of dynamic environments. Construction site observers often need efficient monitoring of workers during construction activities which is still time-consuming and costly to implement. The recent growth of visual data captured actively on construction sites provides different opportunities to enhance autonomous management. This paper introduces an innovative system with a framework for vision-based intelligent scene understanding tailored to construction sites. Our approach leverages real-time object detection, pose estimation, graph-based interaction modeling, and task prediction from visual data simultaneously to analyze sophisticated relationships among workers, tools, and construction activities. The comprehensive assessment with computer vision models realizes brilliant information extraction of construction image and video dataset demonstrating high accuracy in interaction detection and achieving an overall precision of % and a recall rate of %, with real-time processing capabilities at FPS. Furthermore, worker task recognition is well-improved with a spatio-temporal graph-based method using recognized per-frame scene information worker pose, detected bounding boxes of tools, and estimated worker activity. Finally, the case study with visual data assembled from real-world construction projects showed interaction-based scene understanding model achieves % cross-validation accuracy in a framing task, underscoring the potential for this framework to improve efficiency and autonomy in real-time worker-centered construction management.

Keywords: Scene Understanding, Computer Vision, Human-Object Interaction, Construction Sit,

1. Introduction

The growing complexity of construction projects has significantly increased the demands on worker productivity, effective coordination, and task clarity. Modern construction sites are dynamic environments characterized by simultaneous activities, diverse worker roles, and the use of construction tools and machinery. These intricacies necessitate precise management strategies to ensure that tasks are performed efficiently and resources are utilized optimally. However, meeting these demands is increasingly challenging as projects become larger and more complex, requiring managers to monitor multiple processes and workers in real-time. Nowadays, visual data has emerged as a pivotal tool for addressing these challenges [1]. High-resolution cameras, drones, and other imaging technologies can capture a wealth of information about workers, tools, and materials. However, existing applications of computer vision in construction have focused primarily on safety-related tasks, such as identifying violations of personal protective equipment (PPE) or detecting hazardous conditions [2, 3]. Despite the promise of visual data, traditional site monitoring methods remain heavily reliant on manual oversight, which is time-consuming, labor-intensive, and prone to human error. Such potential using computer vision to address comprehensive scene reasoning can enable managers to consistently monitor activities, particularly on large-scale construction sites where resources are dynamic and fast-paced. There is a critical need for solutions that focus on autonomous management that prioritize worker productivity, task allocation, and operational efficiency.

The dynamic nature of construction sites poses significant challenges in understanding worker-centered interactions and predicting construction tasks. Workers often interact with diverse tools and equipment under varying construction conditions with highly variable interaction patterns [4, 5]. These interactions often involve subtle and complex relationships, such as the specific manner in which a worker handles a tool, the sequence of tool usage, or the coordination between workers and workers in shared spaces. Accurately capturing and interpreting these interactions requires robust systems to recognize context-specific nuances. Task prediction in construction is inherently complex due to the sequential and context-dependent nature of activities. Construction tasks may involve multiple workers, tools, and intermediate steps, each of which contributes to the overall process. Developing models that can reliably identify such tasks requires sophisticated feature extraction and contextual dependencies process. Current frameworks for scene understanding fall short in capturing complex and dynamic spatial-temporal relationships inherent in construction scenes. Many approaches rely on static analyses, focusing on isolated frames without considering progression of activities. This limitation

*Corresponding author. Email: @ufl.edu

restricts their ability to model the flow of tasks, which is crucial to understanding construction processes. Furthermore, existing methods treat worker-centered interactions and activity recognition separately rather than understand the scene as interconnected components. This fragmented approach fails to take advantage of the synergistic potential of integrating multiple data modalities. Consequently, these frameworks struggle to provide comprehensive information on task dynamics and worker-centered relationships.

To bridge these gaps, this study introduces an innovative real-time intelligent scene understanding system to address the demands of dynamic construction environments. The proposed system integrates advanced computer vision techniques, including object detection, human pose estimation, and graph-based interaction modeling, to provide insights into worker activities and interactions. Its primary objectives are to accurately model interactions, predict construction tasks dynamically, and provide operational insights in real-time. Leveraging computationally efficient algorithms and optimized pipelines with state-of-the-art detection models, the system identifies workers, tools, and scenes with high precision and minimal latency. Human pose estimation models extract keypoints information, capturing worker movements essential for activity analysis. The system employs a graph-based approach, representing workers and objects as graph nodes and their interactions as edges. Relationships are dynamically inferred from spatial and temporal proximity, object usage patterns, and motion sequence. These intricate dependencies and collaboration patterns can offer a comprehensive understanding of on-site activities. The proposed construction management system can achieve a deeper understanding of interactions and task dynamics, paving the way for smarter construction site operations, empowering project managers to monitor construction activities, anticipate task timelines, and optimize allocation and workflows.

2. Related Works

At present, a considerable amount of research covers a broad range of topics and techniques around computer vision-driven approach in construction industry. Prior research related to this study can be categorized into several categories: object detection, human pose estimation, activity recognition, human-object interaction.

The majority of object detection work focuses on detecting PPE, like hard hat, gloves, and vests, on construction sites. An approach can detect low-level image features like edge of hard hats [6]. Another approach with classifiers, support vector machine (SVM), can detect worker features in video frame [7]. And also, a non-hardhat-use (NHU) detection approach can match human body and hard hat in video frame based on histogram of oriented gradients (HOG) [8]. As the concept of machine learning and deep learning-based detection methods appeared in computer vision and promoted the development of vision-based approach for construction industry to a new pivot. An object detection method with Faster R-CNN model can detect NHU of construction workers which can draw bounding boxes around worker, trained with a dataset of 81,000 annotated image from construction sites [9]. An NHU identification approach with DenseNet-based can realize face detection and bounding box regression [10]. With the introduction of YOLO, an approach can detect the radiation PPE in nuclear power plants in real-time with high accuracy [11]. Another real-time model took advantage of YOLOv3 architecture to detect PPE compliance of workers, verified with CNN-based classifiers like VGG-16, ResNet-50, and Xception [12]. A further Pictor-v2 dataset with 1105 crowdsourcing and 1402 web-mined annotated images was established for YOLO algorithms training to detect common construction objects in real-time [13]. Similarly, Single Shot Detector (SSD) for NHU identification was realized with a dataset GDUT-HWD established to train the SSD-RPA model [14].

Based on the development of computer vision for decades, the opinion of human posture and activity has been promoted and introduced into the construction industry. Even though plenty of pose analysis approached have been provided, the limitation on practical application is still a tough issue, especially in dynamic and complicated construction sites. To estimate human posture and keypoints, a CV approach with Openpose can extract human keypoint and joint information using single camera [15]. Also, a deep learning approach can extract posture like R3DJP with CNN using a single 2D camera, ensuring high accuracy of joints recognition [16]. To better handle video data, approaches have been developed to extract human keypoints and joints information in 2D and 3D pose estimation and visualization [17, 18], which can provide richer data for human activity analysis and recognition [16]. As studies around activity recognition with worker and construction objects have been developed, the initial idea of scene understanding and human-centered interaction were introduced into construction application [19–21]. With deep learning methods, computer vision can recognize features of worker activities in video frames from construction sites. A 2D-vision based network using the combination of CNN and LSTM can accurately recognize worker unsafe activities like climbing ladder from construction site videos [22]. By integrating the idea of the combination of temporal and spatial approach using networks like CNN, YOLO and MDNet, the classification and tracking of worker activities in construction site videos can be realized with better accuracy [23, 24].

Inspired by the prior methods, several studies considered worker activity and tool use by interactive extraction of worker pose, activity, tools, material, and context simultaneously. For example, a further approach introduces probabilistically graphical models to better understand worker activities in far-field surveillance videos [25]. And other studies based on this integrated better detection and analysis models like Faster R-CNN, YOLO and SORT, and C3D for graph-based scene understanding [26, 27]. However, a majority graph-based scene understanding studies focus on safety aspects. An Automated Hazards Identification System (AHIS) can describe construction operation from site videos considering construction safety guidelines but limited to spatial relation [5]. Similarly, safety inspections from site videos can be realized by classifying worker-tool interactions on construction sites based on Faster R-CNN and hand-made rules [28]. Furthermore, another approach in-

troduced Mask R-CNN, graph-based method and C-BERT for interaction-level scene descriptions with construction regulations for hazards inference [29]. A more comprehensive approach for safety monitoring and for hazards identification of construction workers with YOLO and OpenPose for object and human detection for spatial analysis combining a hierarchical scene graph for conditional reasoning [30]. This study builds on the similar inspiration of [30] but differs in objective from safety only inspection to management level construction scene understanding. A spatio-temporal sophisticated method introduced rapid and accurate YOLOv8-based methods for perception and supporting the graph establishment with human keypoint and object as nodes, pairwise interaction as edges, rule-based graph paths as scene description. The approach was presented to realize accurate real-time tool detection, worker pose estimation, construction activity recognition, worker-centered interaction and graph-based scene understanding.

3. Proposed Framework

This section outlines our proposed framework for intelligent scene understanding in industrialized construction, focusing on detecting entities, modeling interactions, and predicting tasks. The framework integrates real-time object detection, graph-based interaction modeling, and temporal task prediction, as illustrated in Figure 1.

3.1. Framework Overview

The framework can be visualized as follows:

- **Input:** Construction site images and videos.
- **Process:** Real-time detection, pose estimation, and graph modeling.
- **Output:** Predicted worker activities and interactions.

3.2. Object Detection and Pose Estimation

The framework includes:

- **Object Detection:** Models to identify tools and workers.
- **Pose Estimation:** Models for extracting keypoints representing worker postures.
- **Implementation Details:** Frameworks such as YOLO for detection and OpenPose for pose estimation.
- **Training Details:** Use of specific loss functions, optimizers, and datasets for effective training.

3.3. Graph-Based Interaction Modeling

3.3.1. Graph Construction

- **Nodes:** Represent workers and tools.
- **Edges:** Represent relationships and interaction metrics.
- **Edge Weights:** Calculated using metrics like Euclidean distance or overlap of bounding boxes.

3.3.2. Spatial-Temporal Dynamics

- Graphs link worker states across frames.
- Edges capture activity progression and task sequences.

3.4. Task Prediction

- **Node Features:** Pose keypoints, object bounding boxes, and activity labels.
- **Edge Features:** Temporal relationships and interaction scores.
- Graph embeddings are processed to classify tasks effectively.

4. Experiment Design

4.1. Dataset Preparation

4.1.1. Data Collection

- Visual data acquisition process.
- Captured construction activities, such as framing.

4.1.2. Annotation Process

- Annotating worker keypoints and poses.
- Labeling bounding boxes for tools.

4.2. Experimental Setup

4.2.1. Hardware and Software

- **Hardware:** GPUs for computational efficiency.
- **Software:** Frameworks like PyTorch and TensorFlow, integrated with real-time processing libraries.

4.2.2. Training and Validation

- Training-validation-test split for robust evaluation.
- Use of augmentation techniques and preprocessing steps.

4.3. Evaluation Metrics

- Precision and recall scores for interaction detection.
- Accuracy for task prediction.
- Frames per second (FPS) for real-time performance.
- Qualitative metrics, such as visualization of results.

5. Results and Discussion

5.1. Quantitative Results

- Interaction detection results, including precision and recall.
- Task prediction accuracy across various activities.
- Real-time processing speed measured in FPS.

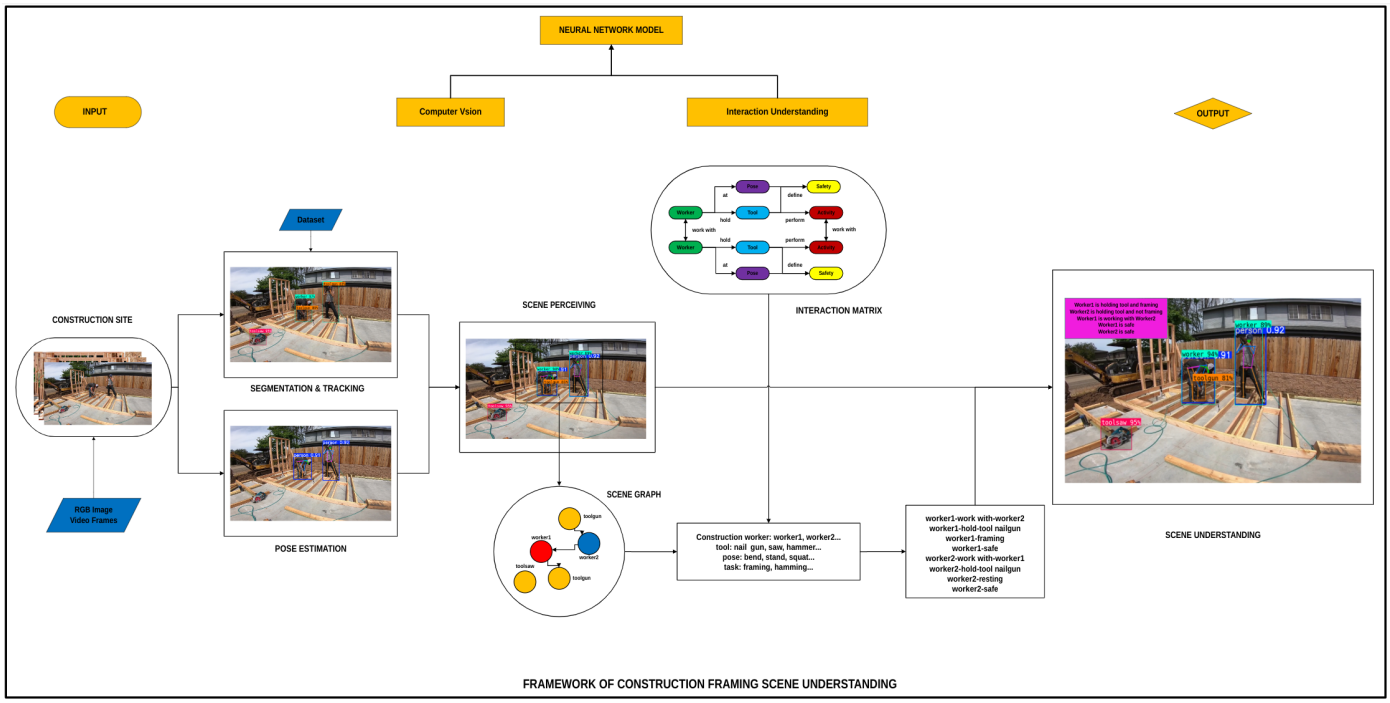

Figure 1: Framework of Construction Framing Scene Understanding.

5.2. *Qualitative Analysis*

- Visualized scenes showing detected interactions and task predictions.
- Interpretability of graph outputs to explain task dynamics.

5.3. *Case Study: Real-World Deployment*

- Analysis of system performance on tasks like framing and assembly.
- Insights into worker-tool interactions and task sequences.

5.4. *Comparison with Baselines*

- Results compared with existing models, highlighting improvements in accuracy and speed.

5.5. *Discussion*

- Strengths and limitations of the proposed system.
- Significance of the framework for industrial applications.

6. Conclusion and Future Work

6.1. *Summary of Contributions*

- Real-time integration of object detection, pose estimation, and graph-based modeling.
- Accurate task prediction using spatio-temporal relationships.

6.2. *Implications for Industrialized Construction*

- Enhancements in task planning, resource allocation, and worker efficiency.

6.3. *Future Directions*

- Incorporating depth data for 3D scene understanding.
- Extending the system for multi-object scenarios and multi-task learning.

References

[1] Kevin K. Han and Mani Golparvar-Fard. “Potential of big visual data and building information modeling for construction performance analytics: An exploratory study”. In: *Automation in Construction* 73 (2017), pp. 184–198. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2016.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0926580516303466.

[2] Mingyuan Zhang, Rui Shi, and Zhen Yang. “A critical review of vision-based occupational health and safety monitoring of construction site workers”. In: *Safety Science* 126 (2020), p. 104658. ISSN: 0925-7535. DOI: https://doi.org/10.1016/j.ssci.2020.104658. URL: https://www.sciencedirect.com/science/article/pii/S0925753520300552.

[3] Weili Fang et al. “Computer vision applications in construction safety assurance”. In: *Automation in Construction* 110 (2020), p. 103013. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2019.103013. URL: https://www.sciencedirect.com/science/article/pii/S0926580519301487.

[4] Shuai Tang, Dominic Roberts, and Mani Golparvar-Fard. “Human-object interaction recognition for automatic construction site safety inspection”. In: *Automation in Construction* 120 (2020), p. 103356. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2020.103356. URL: https://www.sciencedirect.com/science/article/pii/S0926580520309365.

[5] Ruoxin Xiong et al. "Onsite video mining for construction hazards identification with visual relationships". In: *Advanced Engineering Informatics* 42 (2019), p. 100966. ISSN: 1474-0346. DOI: https://doi.org/10.1016/j.aei.2019.100966. URL: https://www.sciencedirect.com/science/article/pii/S1474034619305397.

[6] Kishor Shrestha et al. "Hard-Hat Detection for Construction Safety Visualization". In: *Journal of Construction Engineering* 2015.1 (2015), p. 721380. DOI: https://doi.org/10.1155/2015/721380. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/721380.

[7] Man-Woo Park and Ioannis Brilakis. "Construction worker detection in video frames for initializing vision trackers". In: *Automation in Construction* 28 (2012), pp. 15–25. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2012.06.001. URL: https://www.sciencedirect.com/science/article/pii/S0926580512001136.

[8] Man-Woo Park, Nehad Elsafty, and Zhenhua Zhu. "Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers". In: *Journal of Construction Engineering and Management* 141.9 (2015), p. 04015024. DOI: 10.1061/(ASCE)CO.1943-7862.0000974. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CO.1943-7862.0000974.

[9] Qi Fang et al. "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos". In: *Automation in Construction* 85 (2018), pp. 1–9. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2017.09.018. URL: https://www.sciencedirect.com/science/article/pii/S0926580517304429.

[10] Jie Shen et al. "Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning". In: *Computer-Aided Civil and Infrastructure Engineering* 36.2 (2021), pp. 180–196. DOI: https://doi.org/10.1111/mice.12579. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12579.

[11] Shi Chen and Kazuyuki Demachi. "A Vision-Based Approach for Ensuring Proper Use of Personal Protective Equipment (PPE) in Decommissioning of Fukushima Daiichi Nuclear Power Station". In: *Applied Sciences* 10.15 (2020). ISSN: 2076-3417. DOI: 10.3390/app10155129. URL: https://www.mdpi.com/2076-3417/10/15/5129.

[12] Nipun D. Nath, Amir H. Behzadan, and Stephanie G. Paal. "Deep learning for site safety: Real-time detection of personal protective equipment". In: *Automation in Construction* 112 (2020), p. 103085. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2020.103085. URL: https://www.sciencedirect.com/science/article/pii/S0926580519308325.

[13] Nipun D. Nath and Amir H. Behzadan. "Deep Convolutional Networks for Construction Object Detection Under Different Visual Conditions". In: *Frontiers in Built Environment* 6 (2020). ISSN: 2297-3362. DOI: 10.3389/fbuil.2020.00097. URL: https://www.frontiersin.org/journals/built-environment/articles/10.3389/fbuil.2020.00097.

[14] Jixiu Wu et al. "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset". In: *Automation in Construction* 106 (2019), p. 102894. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2019.102894. URL: https://www.sciencedirect.com/science/article/pii/S092658051930264X.

[15] Manlio MassirisFernández et al. "Ergonomic risk assessment based on computer vision and machine learning". In: *Computers Industrial Engineering* 149 (2020), p. 106816. ISSN: 0360-8352. DOI: https://doi.org/10.1016/j.cie.2020.106816. URL: https://www.sciencedirect.com/science/article/pii/S0360835220305192.

[16] Hong Zhang, Xuzhong Yan, and Heng Li. "Ergonomic posture recognition using 3D view-invariant features from single ordinary camera". In: *Automation in Construction* 94 (2018), pp. 1–10. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2018.05.033. URL: https://www.sciencedirect.com/science/article/pii/S0926580518302231.

[17] Yantao Yu et al. "Joint-Level Vision-Based Ergonomic Assessment Tool for Construction Workers". In: *Journal of Construction Engineering and Management* 145.5 (2019), p. 04019025. DOI: 10.1061/(ASCE)CO.1943-7862.0001647. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CO.1943-7862.0001647.

[18] Wenjing Chu et al. "Monocular Vision–Based Framework for Biomechanical Analysis or Ergonomic Posture Assessment in Modular Construction". In: *Journal of Computing in Civil Engineering* 34.4 (2020), p. 04020018. DOI: 10.1061/(ASCE)CP.1943-5487.0000897. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000897.

[19] SangUk Han, SangHyun Lee, and Feniosky Peña-Mora. "Vision-Based Detection of Unsafe Actions of a Construction Worker: Case Study of Ladder Climbing". In: *Journal of Computing in Civil Engineering* 27.6 (2013), pp. 635–644. DOI: 10.1061/(ASCE)CP.1943-5487.0000279. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000279.

[20] Ardalan Khosrowpour, Juan Carlos Niebles, and Mani Golparvar-Fard. "Vision-based workface assessment using depth images for activity analysis of interior construction operations". In: *Automation in Construction* 48 (2014), pp. 74–87. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2014.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0926580514001824.

[21] Jiannan Cai, Yuxi Zhang, and Hubo Cai. "Two-step long short-term memory method for identifying construction activities through positional and attentional cues". In: *Automation in Construction* 106 (2019), p. 102886. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2019.102886. URL: https://www.sciencedirect.com/science/article/pii/S0926580519302316.

[22] Lieyun Ding et al. "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory". In: *Automation in Construction* 86 (2018), pp. 118–124. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2017.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0926580517302650.

[23] Xiaochun Luo et al. "Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks". In: *Automation in Construction* 94 (2018), pp. 360–370. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2018.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0926580517311019.

[24] Xiaochun Luo et al. "Vision-based detection and visualization of dynamic workspaces". In: *Automation in Construction* 104 (2019), pp. 1–13. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2019.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0926580518312706.

[25] Xiaochun Luo et al. "Capturing and Understanding Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning". In: *Computer-Aided Civil and Infrastructure Engineering* 34.4 (2019), pp. 333–351. DOI: https://doi.org/10.1111/mice.12419. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12419.

[26] Xiaochun Luo et al. "Recognizing Diverse Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks". In: *Journal of Computing in Civil Engineering* 32.3 (2018), p. 04018012. DOI: 10.1061/(ASCE)CP.1943-5487.0000756. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000756.

[27] Xiaochun Luo et al. "Combining deep features and activity context to improve recognition of activities of workers in groups". In: *Computer-Aided Civil and Infrastructure Engineering* 35.9 (2020), pp. 965–978. DOI: https://doi.org/10.1111/mice.12538. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12538.

[28] Shuai Tang and Mani Golparvar-Fard. "Machine Learning-Based Risk Analysis for Construction Worker Safety from Ubiquitous Site Photos and Videos". In: *Journal of Computing in Civil Engineering* 35.6 (2021), p. 04021020. DOI: 10.1061/(ASCE)CP.1943-5487.0000979. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000979.

[29] Lite Zhang et al. "Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge". In: *Automation in Construction* 142 (2022), p. 104535. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2022.104535. URL: https://www.sciencedirect.com/science/article/pii/S092658052200406X.

[30] Shi Chen and Kazuyuki Demachi. "Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph". In: *Automation in Construction* 125 (2021), p. 103619. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2021.103619. URL: https://www.sciencedirect.com/science/article/pii/S0926580521000704.