A Real-time Intelligent Scene Understanding System for Human-Centered Dynamic and Interactive Construction Tasks

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 % accuracy, % higher than popular current approach, in a framing task which can be generalized by applying different task knowledge, 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. Methodology

The proposed method of vision-based real-time intelligent scene understanding for human-centered dynamic and interactive construction task is outlined in detail in this section, focusing on detecting entities, modeling interactions, and predicting tasks. The framework integrates real-time object detection, pose estimation, graph-based interaction modeling, and spatio-temporal task prediction, as illustrated in Figure 1. The autonomous system is working with a complicated sequence starting with the real-time RGBD image and video frames as input. The system is then forward to entity detection for construction tools usage based on the advanced YOLOv8 [31]. At the same time the tracking of individual workers focusing on the keypoints based on YOLOv8-pose was integrated to finish the perceiving. Then with all information achieved, simultaneous analysis can be performed through introducing sophisticated relationships among workers, tools, and construction activities. The relationships including worker-worker, worker-tool are associated with interaction matrixes tailored to construction tasks which can be visualized as a graph with relationship between individuals. Based on the extraction and association, the understanding of the scene can be achieved with the spatial relationships to refine the graph to describe pairwise interaction. Furthermore, the temporal information can be integrated to gather the sequence of all interactions and predict the dynamic construction tasks, realizing autonomous whole construction site management.

3.1. Scene Perceiving

3.1.1. Tool Detection

The detection and localization of construction tools is performed with computer vision based object detection deep learning model. The majority of modern models can be generally divided into two categories: two-stage approaches and one-stage approaches. Two-stage approaches, such as the R-CNN family, operate in two steps: initially proposing regions of interest through selective search or regional proposal networks, and subsequently refining these regions with a classifier. In contrast, one-stage approaches, such as YOLO firstly introduced in 2016 as a game changer and its variants gradually promoting the boundary of computer vision until YOLOv11 [31, 32], eliminate the region proposal stage and directly run detection over a dense sampling of possible locations, enabling simpler architectures and faster inference. The YOLO family achieves high precision and recall, ensuring robust real-time detection in varying and dynamic construction sites.

YOLOv8, supported by the same core as YOLOv11, represents a significant evolution in the YOLO family, building upon its predecessors such as a more streamlined backbone network and an advanced detection head.with improved accuracy, scalability, and performance [31]. Unlike earlier versions, YOLOv8 introduces an adaptive anchor-free detection mechanism, advanced network designs, and data augmentation techniques that enhance its ability to detect small and densely packed objects, ideal for construction tools in dynamic and cluttered environments. The YOLOv8 architecture is established with three integral components: the backbone, neck, and head. The backbone, a modified version of CSPDarknet53, extracts features from the input images, capturing both spatial and semantic information. The neck incorporates a combination of a feature pyramid network (FPN) and a path aggregation network (PAN) to improve semantic expression and localization across multiple scales and resolutions. Finally, the head employs an anchor-free detection model with a decoupled head to independently predict bounding box positions and object classes.2

YOLOv8 uses loss functions to optimize classification and bounding box loss including binary class entropy loss (BCEL), distribution focal loss (DFL), and CIoU loss function: Binary Cross-Entropy Loss (BCEL) for classification:

$$L_{\text{class}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)),$$

where y_i is the ground truth class, p_i is the predicted probability, and N is the total number of samples. Distribution Focal Loss (DFL) for bounding box regression:

$$L_{\text{DFL}} = \sum_{i=1}^{N} \text{Softmax}(t_i) \cdot \log(\hat{t}_i),$$

where t_i and $\hat{t}_i$ are the predicted and target distributions for bounding box localization. Complete IoU (CIoU) Loss for bounding box refinement:

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^g)}{c^2} + \alpha v,$$

where $\rho(b, b^g)$ is the Euclidean distance between the center of the predicted box b and the ground truth box b^g, c is the diagonal length of the smallest enclosing box, α is a trade-off parameter, and v measures the aspect ratio consistency.

The final loss function is a weighted sum of the above components:

$$L = \lambda_{\text{class}} L_{\text{class}} + \lambda_{\text{DFL}} L_{\text{DFL}} + \lambda_{\text{CIoU}} L_{\text{CIoU}},$$

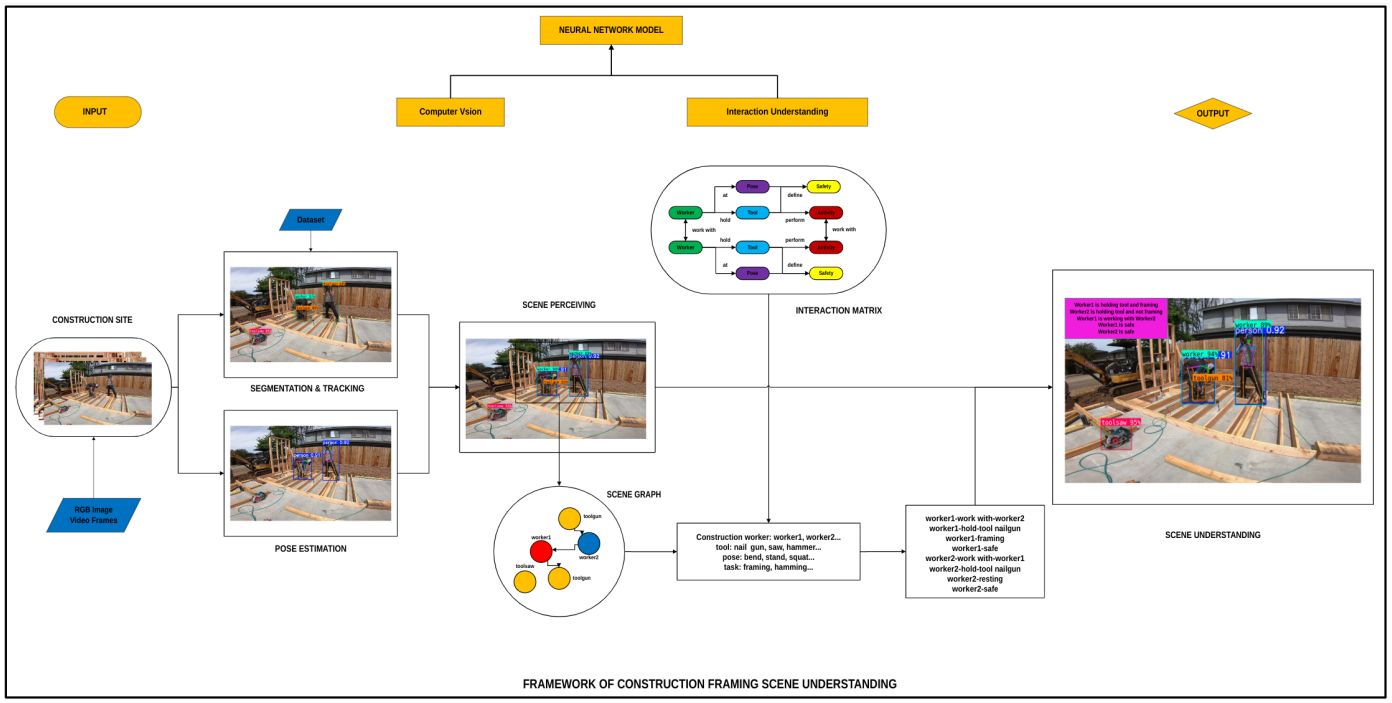

Figure 1: Framework of Construction Scene Understanding.

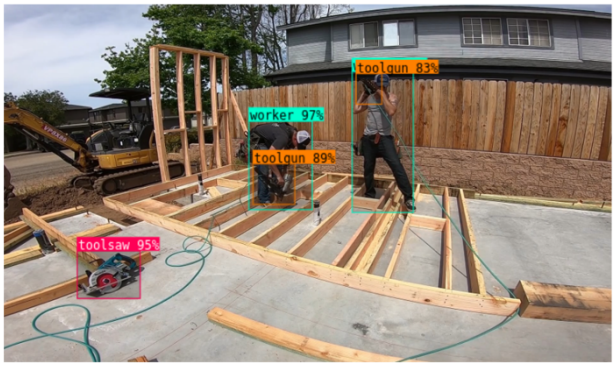

Figure 2: YOLOv8-based tool detection.

where $\lambda_{\text{class}}, \lambda_{\text{DFL}}, \lambda_{\text{CIoU}}$ are the weights assigned to each loss component.

Additionally, YOLOv8 enhances its robustness through online data augmentation during training. This augmentation process introduces variations in each epoch to improve generalization. HSV Augmentation: Adjusts hue, saturation, and value for color diversity. Image Translation: Shifts objects spatially within the image. Image Scaling: Alters the scale of objects to accommodate size variability. Image Flipping: Mirrors images horizontally or vertically to account for orientation changes. Mosaic Augmentation: Combines four images into one, providing diverse object arrangements for complex scenes.

The combination of advanced architecture, efficient loss functions, and robust data augmentation make YOLOv8 a robust solution for construction tool detection. It excels in identifying tools such as hammers, drills, and nailer in dynamic, cluttered construction site environments for interaction modeling.

3.1.2. Worker Pose Estimation

Inspired by tool detection, construction worker pose estimation is achieved using YOLOv8-Pose, an extension of YOLOv8 tailored for human keypoints detection, representing the latest advancements in real-time technology. YOLOv8-Pose builds upon the YOLOv8 architecture by incorporating pose estimation capabilities, making it highly efficient for real-time applications in construction site environments, expanding the application scope of the model. The model then applies a multi-scale approach to detect bounding boxes for individuals and simultaneously predict keypoints for each person. These keypoints correspond to 17 body joints, including shoulders, elbows, hips, knees, and ankles to determine worker posture and movements, as defined by the COCO Keypoints Challenge dataset. This keypoints are critical for understanding worker activities and their interactions with tools. The pose data is subsequently integrated into the interaction modeling process to provide a comprehensive understanding of the scene. 3

The structure of YOLOv8-pose is based on YOLOv8 which can be divided into three main parts, the backbone, neck and head networks, optimized to ensure accuracy and reliability of the model recognition results under dynamic construction sites. With pre-processing of the raw image data, the input data is standardized for model to ensure stability. The backbone is still the core for feature extraction using convolutional layers and advanced modules C2f, ensuring efficiency . The neck network can fuse features at various scales using up sampling and concatenation layers, integrating spatial and semantic information to improve accuracy. The head is adapted for pose estimation

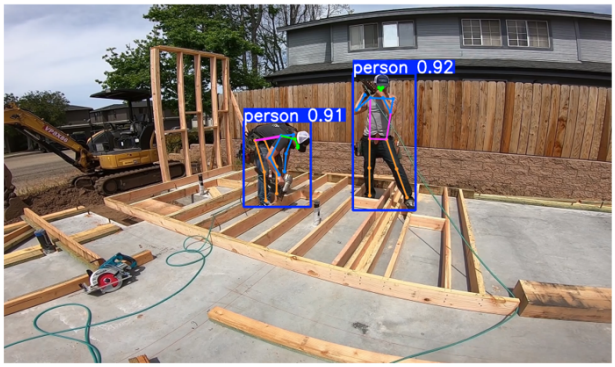

Figure 3: YOLOv8-Pose-based worker pose estimation.

to predict simultaneously with outputs for bounding box coordinates and keypoints coordinates in real-time object. YOLOv8-Pose also adopts an anchor-free design to simplify the detection process and improve computation and inference speed. The method has ultimate performance particularly for real-time applications such as worker activity recognition in challenging construction environments.

The loss function for YOLOv8-Pose comprises two major components, compared to YOLOv8 including BCE, DFL, CIoU, the loss function of keypoints and keypoints confidence is added to total loss function:

$$L_{\text{total}} = \lambda_{\text{det}} L_{\text{det}} + \lambda_{\text{key}} L_{\text{key}},$$

where λ_{det} and λ_{key} are weights balancing the contributions of the keypoint and bounding box losses.

The efficient loss functions and robust data augmentation after end-to-end training can provide YOLOv8-Pose high efficiency and real-time inference, making it a robust and real-time solution for construction worker pose estimation for precise data for interaction prediction.

3.2. Scene Understanding

3.2.1. Worker-Centered Scene Graph Establishment

With the scene information extracted with tool detection and worker pose estimation, the scene graph can be established with the knowledge-based matrix integration. To start with the graph, workers, kyepoints of worker, and construction tools can be the potential nodes to be connected. In this framework, we use $W = \{W_1, W_2, \ldots, W_N\}$ represent the workers detected in a scene, where N is the number of workers. However, the keypoints $W_n = \{(x_n^{(0)}, y_n^{(0)}), (x_n^{(1)}, y_n^{(1)}), \ldots, (x_n^{(17)}, y_n^{(17)}), id_n\}$of the worker obtained using YOLOv8-Pose will support the nodes as information as well as worker-id. As each worker's pose is defined by the set of body keypoints and the activities can be infered from the pose. Similarly, we use $T = \{T_1, T_2, \ldots, T_M\}$ represent the set of detected construction tools, where M is the number of detected objects. And the information for all nodes of construction tool can be enriched by the bounding box coordinates detected by YOLOv8 including the center of the bounding box and width and height $T_m = (x_m, y_m, w_m, h_m, c_m, cl_m, lb_m, id_m)$ as well as the object class, label and tool-id. **??**

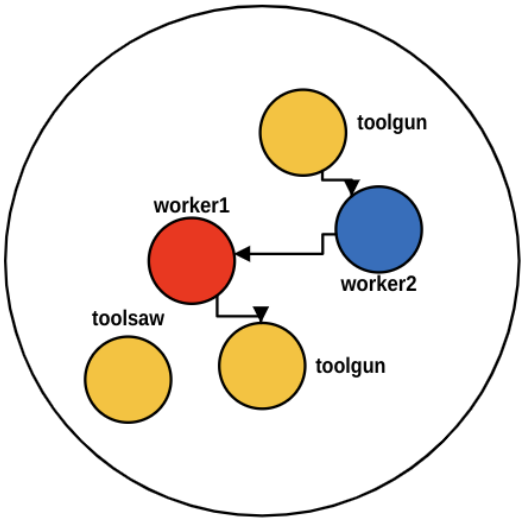

Figure 4: Worker-Centered Scene Graph Establishment

To associate workers with other workers and tools, we use the relationships between the keypoints and tool bounding boxs for example, the keypoints of both hands of worker i are within the area of the bounding box of tool j, the geometric center of shoulders and hips of worker i and j within the distance threshold. Based on this, we construct a graph $G = (U \cup V, E)$, where $U = \{1, 2, \ldots, N\}$ represents the set of workers, $V = \{1, 2, \ldots, M\}$ represents the set of tools, and E is the set of relationship as edge. The weight of an edge $e = (u, v) \cup (u, u) \in E$ is determined by the Euclidean distance between the detected key parameters of a worker W_u and the bounding box center of an object T_v:

$$w(u, v) = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2},$$

where (x_u, y_u) is the location of the detected hand keypoint of worker W_u, and (x_v, y_v) is the center of the bounding box for object T_v. Similarly, the Euclidean distance between the detected and calculated key centers of a worker W_{u1} and another worker W_{u2}:

$$w(u1, u2) = \sqrt{(x_{u1} - x_{u2})^2 + (y_{u1} - y_{u2})^2},$$

where (x_u, y_u) is the location of the detected and calculated geometric center of worker W_u.

Furthermore, based on the simple pairwise relationship, we introduced a completely new idea, adding an edge from worker i back to worker i based on the activity-focused analysis using joint angles calculated from trio consequent keypoints $(x_u^{i,j,k}, y_u^{i,j,k})$:

$$w(u1, u1) = \arccos\left(\frac{((x_i - x_j), (y_i - y_j)) \cdot ((x_k - x_j), (y_k - y_j))}{\| (x_i - x_j, y_i - y_j) \| \cdot \| (x_k - x_j, y_k - y_j) \|}\right),$$

where the numerator represents the dot product of the two vectors and the denominator is the product of the Euclidean norms of the vectors. This weight of the joint angle is used to analyze posture-based construction activities and tasks.

3.2.2. Knowledge-Based Scene Understanding

Advancements in graph-structured methods have enabled structured representations of visual scenes. Building upon the concept of scene graphs, we propose a novel scene graph with comprehensive nodes for detected tools, workers and keypoints, advanced pairwise relationships between worker-worker, worker-tool, spatial worker activity edges to model worker-centered interactions and activity reasoning in dynamic construction environments. However, this simple scene graph cannot provide complex construction scene understanding for example, nailer must be used when nailing and only spending a period of nailing posture can be inferred as a framing task. Considering this, we propose a novel knowledge-based matrix into scene graph to perform high-level scene understanding for construction management. 5

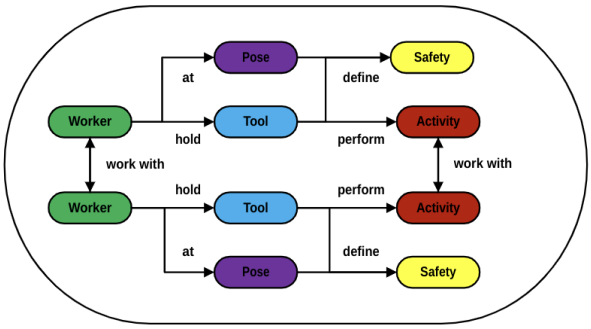

Figure 5: Worker-Centered Relationship Matrix.

To ensure the application and potential, the framework is focusing generalize all kinds of construction tasks with worker involved, tool usage, tool type, and potential posture, which can be introduce to the system as a source to refine the graph-based approach. We first introduce the idea of triplet based on pairwise relationships to better describe the construction scene. For task prediction, the triplet can be "worker1 holding tool nail gun and performing framing task", which will include "node worker" embedded id1, "node tool" embedded label nail gun, "edge worker1 to tool nail gun" embedded hold, and "edge worker1 to worker1" embedded framing task. This will provide a comprehensive scene description and understanding. 6

Moreover, with the introduction of temporal approach, the edge representing activity can be better refined when considering the sequence of posture to meet the confidence threshold for activity recognition. The frame information at time t can be fed to the next frame at time t+1 and the memory of a period T of real-time frame scene graph can be store to robust and stabilize the understanding. Combining with robust spatial relationship used for interaction prediction, the whole spatio-temporal scene understanding can achieve ultimate performance for real-time dynamic construction tasks management and enforcement. 7

4. Experiment and Results

4.1. Experimental Setup

Data was collected from real-world construction sites, capturing diverse activities and interactions. The dataset was annotated with worker poses, tool locations, and task labels.

4.2. Evaluation Metrics

The system performance was evaluated using precision, recall, F1-score for detection tasks, and accuracy for task prediction. Real-time performance was measured in frames per second (FPS).

4.3. Results

The framework achieved high detection accuracy, with precision and recall exceeding % for most categories. Pose estimation reliably identified keypoints, enabling accurate interaction modeling. Task prediction using ST-GNN demonstrated % accuracy, showcasing the method's effectiveness.

5. Discussion

5.1. Strengths

The proposed framework effectively integrates multiple computer vision techniques, achieving high accuracy in real-time scenarios. The use of dynamic graphs enhances interaction modeling, improving task prediction.

5.2. Limitations

The system's performance depends on the quality of training data. Future work should address the variability of construction site conditions to improve generalizability.

5.3. Future Directions

Expanding the dataset and incorporating advanced graph reasoning methods could further enhance the framework's capabilities. Exploring additional applications, such as worker safety monitoring, is another promising direction.

6. Conclusion

This study introduces a novel real-time intelligent scene understanding system for construction site management. By integrating object detection, pose estimation, and graph-based reasoning, the framework addresses the challenges of dynamic construction environments. The results demonstrate its potential for improving efficiency and task understanding. Future work will focus on expanding applications and improving model generalization.

Acknowledgements

The authors thank the University of Florida for support and access to construction site data for this study.

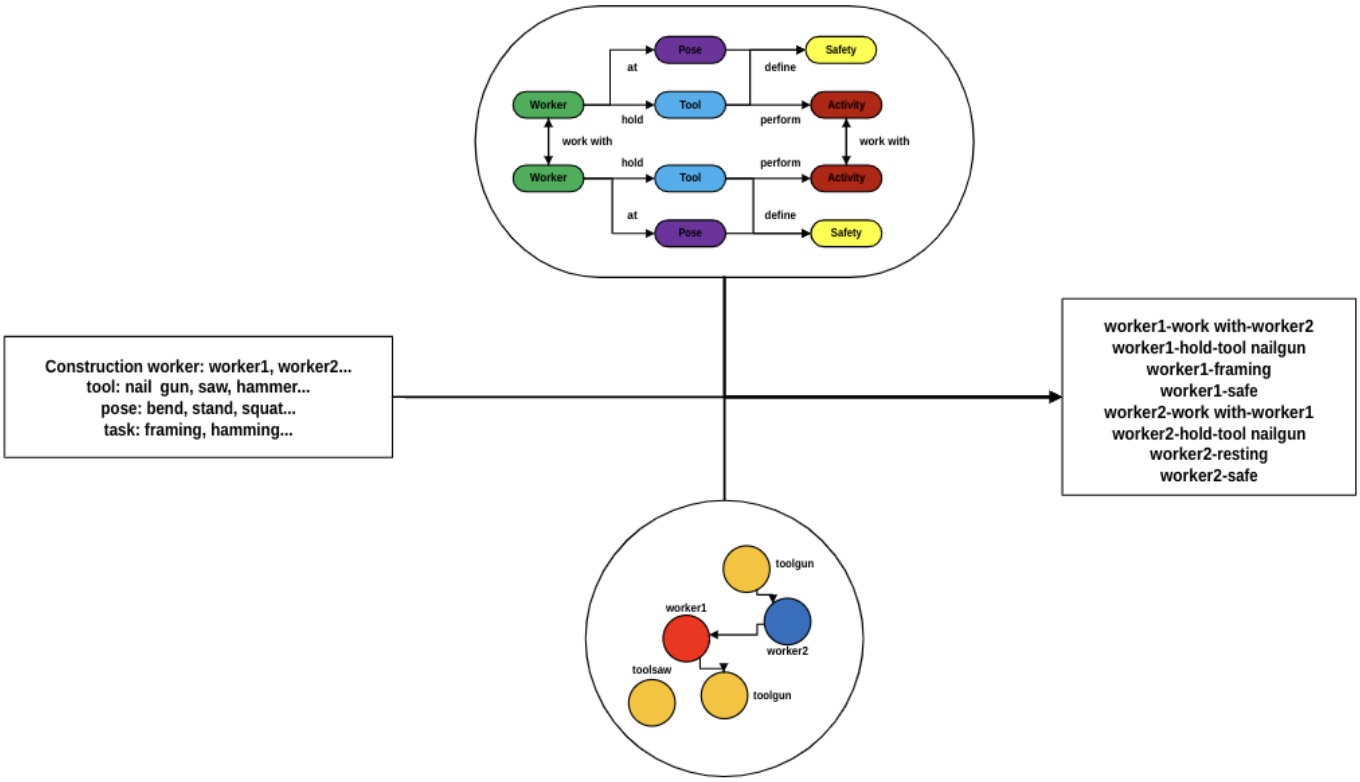

Figure 6: Knowledge-Based Construction Description.

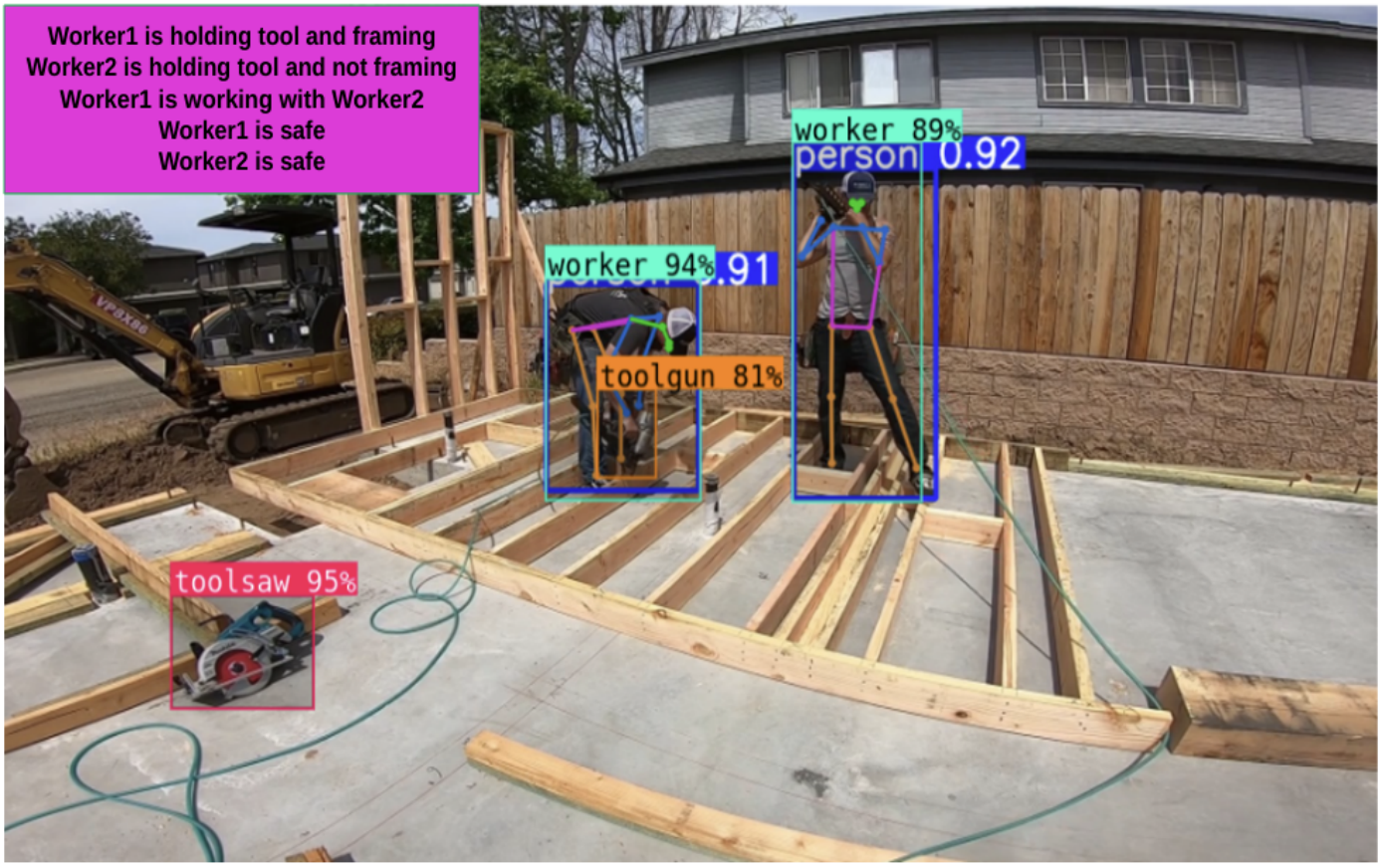

Figure 7: Scene Understanding on Construction Site.

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.

[31] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. *Ultralytics YOLO*. Version 8.0.0. Jan. 2023. URL: https://github.com/ultralytics/ultralytics.

[32] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.