Video-Based Productivity Monitoring of Worker and Large-Scale Object Interactions in Construction Sites

Mik Wanul Khosiin[1], Jacob J. Lin[1], Eko Andi Suryo[2], Kartika Puspa Negara[2], Ismiarta Aknuranda[3], and Chuin-Shan Chen[1]

[1]Department of Civil Engineering, National Taiwan University, Taipei City, Taiwan
[2]Department of Civil Engineering, Universitas Brawijaya, Malang City, Indonesia
[3]Department of Information Systems, Universitas Brawijaya, Malang City, Indonesia

mikwk@caece.net, jacoblin@ntu.edu.tw, dchen@ntu.edu.tw

Abstract -

Automating productivity monitoring is crucial for improving the construction industry. To measure productivity, we should identify which worker works for what object and their relationship. The lack of understanding between human and object interaction in a large-scale format from video surveillance has become a significant challenge in construction sites. However, the existing vision-based studies only focus on object detection and activity recognition, which do not recognize workers, objects, and actions simultaneously. This situation makes managers unable to measure the productivity of the workers effectively. To address the issue, this study applies the HOI technique, which consists of object detection tasks and interaction prediction tasks through faster R-CNN and graph neural networks (GNN). There are two groups of actions in this interaction, including productive (installing, preparing, and transporting) and non-productive actions (no interaction) on the formwork structure. Our model achieves 0.674, 0.556, and 0.632 mAP scores of the local area, global area, and average area of the objects sequentially, indicating that the model can monitor construction productivity effectively. For future studies, utilizing more information, such as temporal and body postures of workers, can potentially improve the performance of the HOI model for the productivity monitoring process.

Keywords -

Productivity monitoring; human-object interaction; graph neural networks (GNN)

1 Introduction

The construction industry is less competitive than other sectors. From 2000 to 2022, It only grew 0.4 percent annually and has a 1.6 percent and 2,6 percent gap towards the world total economy and manufacturing industry simultaneously ([1]). Even though global productivity was boosted by China's improvement (4 percent) in 2021, it shows signs of slowing down due to the ineffective workforce monitoring process. The main issue is that managers cannot improve their performance because they lack understanding of the interaction between workers, objects, and their actions ([2]) simultaneously. This situation brings practical problems, such as rework activities, defective products, construction accidents, etc. Moreover, the current technology uptake is mainly focused on controlling systems rather than productivity. This means a comprehensive investigation of workers is essential to help the managers in the decision-making process.

Currently, to understand the interactions between humans and objects in construction sites, managers must conduct real observations consistently ([3], [4]), which is time-consuming and low accuracy. The current technologies applications in construction sites only focus on detection and activity recognition, and their interaction has not yet been addressed ([5]). This means the prompt about which workers work for what objects remains unsolved and becomes a crucial issue in the productivity monitoring system ([6]). Moreover, construction sites have abundant data collected by managers for daily reports from the site's video surveillance. This group of data is dominated by large-scale instances for both humans and objects. However, this data is only used for administrative purposes and has not been organized and utilized optimally for the productivity monitoring process. Thus, the potential technology is required to explore this data and generate valuable information about worker activity.

Human-object interaction (HOI) [7] represents a transformative vision-based methodology that enables simultaneous detection and analysis of humans, objects, and their interactions within an image or video. This technique is particularly valuable on construction sites, where monitoring worker productivity and interactions with various elements is crucial. By employing HOI models, supervisors can gain a detailed understanding of how workers engage with tools and materials, which can significantly enhance operational efficiency and safety protocols. The ability to identify and categorize these interactions in real time offers numerous benefits. It allows for immediate responses to safety hazards or inefficiencies, and the data collected can be used to refine workflow designs and worker training programs. Moreover, integrating HOI technology into existing surveillance systems transforms passive monitoring into an active management tool. This facilitates a more proactive approach to project management and quality control, ensuring that all activities on site align with safety and productivity standards. Through the precise detection of interactions, HOI technology not only boosts site management but also contributes to the development of smarter, safer construction environments.

To address this challenge, this study proposes the HOI utilizing graph theory in large-scale instances to understand the connection between workers and their elements. This study consists of five sections: first, the Introduction presents the topic and its importance, while Related Studies reviews previous work and contextualizes the research. The Methodology section explains the approaches and techniques used for HOI detection. The experiment details the dataset preparation and results of practical tests conducted. Finally, the Conclusion summarizes the findings and suggests future research directions.

2 Related Studies

To understand the process of the HOI technique ([8], [9], [9], and [10]), we need to compare it with the activity recognition task. In the activity recognition task, there are only two steps, including object detection and activity prediction. Object detection in this task only localizes the humans and uses their spatial and temporal features to determine the type of action. In Figure 6, the activity recognition can identify the worker doing the installing action on the particular object. However, without detecting the objects and generating the interaction with the humans, activity recognition can not address a practical issue in construction sites that requires a comprehensive understanding of workers for the productivity monitoring process.

While in the HOI task, this technique offers more complete detection information, such as humans, objects, connection, and action prediction. To obtain those outputs, the HOI task requires three steps, consisting of object detection, interaction detection, and activity prediction, where this approach has one more additional step than the activity recognition task. After evaluating both tasks, we can conclude that there are two important processes, including object detection and activity prediction. To have an accurate HOI prediction, four elements from the frames must be extracted: humans, objects, interaction, and actions with graph neural networks (GNN)([11]). Thus, we summarize the recent construction worker and objects-related studies in Table 1 below that utilized several approaches to highlight the research gap in this study. In addition, the two important tasks will also be elaborated in this section to highlight the technical gap in this study.

Table 1. The recent studies in monitoring construction workers

Authors	Worker monitoring	Limitations
Q. Zhang ([12])	object detection	no obj., act., and int.
Y. Sun ([13])	object detection	object only
R. Mutegeki ([14])	activity recognition (CNN-LSTM)	no obj., and int.
S. Garcia ([15])	activity recognition (YOLO)	no obj., and int.
M.-Y. Cheng ([16])	action recognition (YOWO)	no obj., and int.
S. Tang ([17])	HOI for safety	safety only

Object detection offers classification and localization tasks. We can identify the type and position of the objects in each frame by using several existing models, such as You Only Look Once (YOLO) ([18]), Faster R-CNN ([19]), Mask R-CNN ([20]), and others. In the current construction practices, Q. Zhang ([12]) detects the construction workers and classifies them based on the color of the hat using Faster R-CNN. However, this study has not considered the materials around the workers and the type of action of the workers, so this study provides only one HOI element. Moreover, Y. Sun ([13]) offers to detect construction materials, such as concrete, brick, metal, wood, and stone, through the VGG16 model. However, they focus solely on the objects and ignore the other two important elements in HOI tasks (humans and actions). Therefore, the object detection task is not enough to fulfill the required information in construction practices. We need to utilize potential techniques to enhance the recognition level in construction sites.

Activity recognition is the deep learning task that mainly focuses on understanding the type of action humans take. The current models can help us conduct action recognition consisting of Long Short-Term Memory (LSTM) ([21]), YOLO, and You Only Watch Once (YOWO) ([22]), and etc. Some studies have implemented these techniques in construction sites, like R. Mutegeki ([14]) recognizing the worker action by combining Convolution Neural Networks (CNN) ([23]) and LSTM models. Furthermore, S. Garcia ([15]) and M.-Y. Cheng ([16]) observed the worker activity by using YOLO and YOWO sequentially. However, Those studies did not provide information about objects and their interaction in construction sites. Consequently, construction practitioners cannot fully understand what materials are being used or processed by which workers.

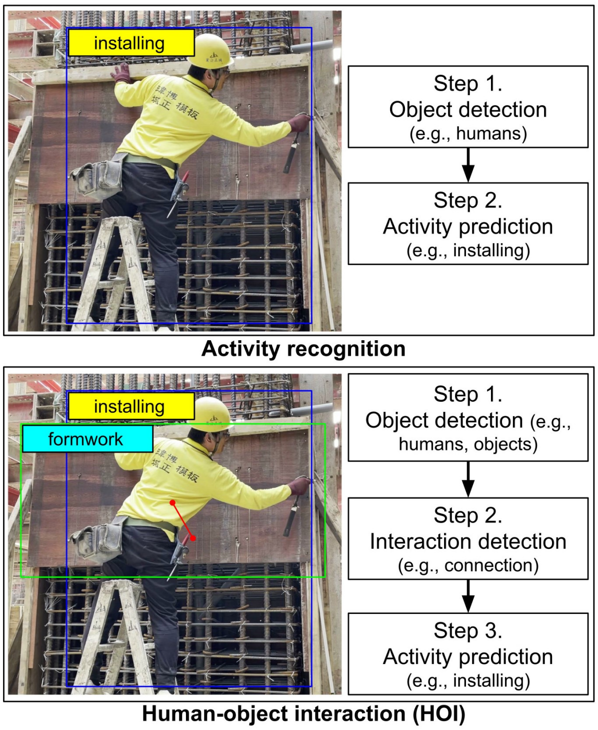

Figure 1. Activity recognition and HOI task comparison

This incomplete situation triggers us to find a model that can provide comprehensive information about workers.

From Table 1 and In Figure 6, we can imply that the current studies lack interaction information among the workers. Moreover, the objects located around workers have not been completely recognized, which adds more challenges. Therefore, identifying humans, objects, actions, and their interaction as a single information is an essential issue to enhance the construction monitoring process.

3 Methodology

Based on the previous discussion, this study also has three steps (Figure 2) in utilizing the HOI technique for productivity monitoring in construction sites, such as Step 1. Object detection, and Step 2. Interaction prediction with GNN, and Step 3. Activity recognition. In Step 1, we utilize Faster R-CNN ([19]) to localize the objects and extract potential features, such as visual features, semantic features, and spatial features. Then, these features will be stored in the fully connected layers (fc7) as input for the GNN ([11]) in the next step. In Figure 2, W1, W2, O1, and O2 stand for worker-1, worker-2, object-1, and object-2, which represent the detection results that consist of two HOI elements, humans and objects. These results will be considered as the node data, the initial information used to build the edge between one node and another. Furthermore, despite using a large-scale format, we still consider many to many detections, which means we can potentially identify crowd humans and objects in each frame. From those two data, we can generate multiple interactions to obtain a better understanding of the worker's activities. In other words, if we detect fewer people and objects in the image, we will also have less interaction prediction, or it relies on the performance of the object detection task. Finally, object detection is crucial in HOI; the higher the object detection performance, the better the productivity monitoring results will be achieved.

In Step 2, the features from the previous step will be processed in the GNN. We identify the potential interaction or edge candidates (red arrow) through attention layers to update the features in each node and construct the final graph structures. We utilize the graph attention networks (GAT) ([24], [25]) as one of the variations of GNN structure that offers a comprehensive node's analysis. Then, we follow these mathematical operations (Eq. 1 and Eq. 2) as the GAT concept to prepare the HOI model.

$$G = (V, E, \{h'_1, h'_2, h'_3, \ldots, h'_n\}) \tag{1}$$

$$h'_i = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j\right) \tag{2}$$

Where G is defined as a graph, V represents vertices that consist of $v_1, v_2, \ldots, v_n$, and E stands for the edge of the graph $\{w_{ij} \in V\}$. Before we build the Graph, we should calculate h'_i as the updated feature vectors for each node i, obtained by applying a non-linear activation function σ (such as the ReLU function) to a weighted sum of the transformed feature vectors Wh_j of neighboring nodes j in the set $N(i)$. The weights α_{ij} are attention coefficients obtained from the training process through the attention networks that determine the importance of each neighbor's features to node i, computed by the network based on the nodes' relative positions or feature similarities. Once we have the updated node vector h'_i, we can combine the whole features and generate the final graph structure for the prediction process. In addition, since we most likely detect many workers and objects in one frame, this study also proposes the multiple-heterogeneous interaction prediction. Ultimately, we will gain a thorough understanding of the intricate interactions between workers and the objects they handle on construction sites.

In Step 3 of the workflow, the activity recognition utilizes enriched graph data from GNNs to predict specific interactions like "installing" or "transporting" between workers and objects on construction sites. This stage critically assesses interaction dynamics using contextual and relational data of the nodes, which represent workers and

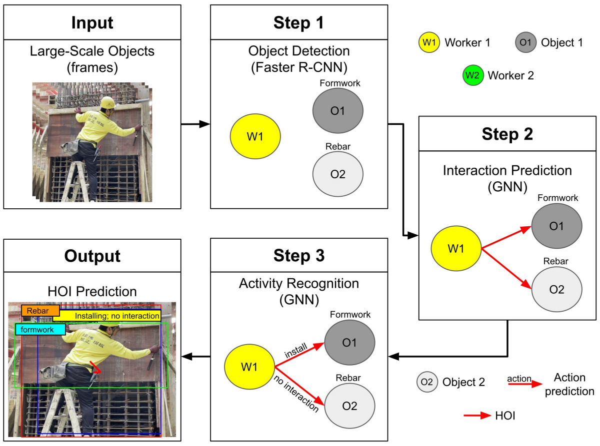

Figure 2. General workflow of productivity monitoring using HOI technique

objects. The GNN refines features and attention weights to highlight significant interactions based on historical and spatial-temporal relationships, providing supervisors with insights for resource allocation and workflow optimization. For evaluation, the model's performance is quantified using the mean Average Precision (mAP) score, which assesses the precision of the predicted interactions across various confidence thresholds. This metric helps ensure the accuracy and reliability of the interaction predictions, supporting effective decision-making in productivity monitoring.

In conclusion, the implementation of HOI techniques using GNNs offers a transformative approach to productivity monitoring on construction sites. This research, structured in three strategic steps: object detection, interaction prediction, and activity recognition, enables precise monitoring and analysis of worker interactions with site elements. Applying GNNs facilitates real-time processing and complex interaction analyses, significantly improving operational efficiency and safety management. The adaptability of this methodology to diverse construction environments underscores its potential to revolutionize project management strategies. By leveraging such advanced technologies, construction managers can optimize workflows, enhance safety protocols, and improve overall resource allocation. Continued advancements and integration of broader data types and interaction scenarios could further elevate the effectiveness of this model, setting new benchmarks in technology-driven construction management and enhancing both efficiency and safety on sites.

4 Experiment

We utilize the supplementary material from M.-Y. Cheng, 2023 [16] provided a set of figures about construction workers who work in installing formwork structures in a large-scale and high-resolution format. We collected 1250 images and divided them into training (1000 images) and testing sets (250 images). We use the Hico-det-UI ([26]) for the labeling process, and we provide two labels (bounding boxes and the action ID) and three classes, which consist of worker, formwork, and steel rebar. Moreover, since those materials are installed in large locations in construction sites, we implement local and global areas that refer to the associated area of objects with direct interaction with the workers. Then, we prepare a list of potential actions, such as preparing and installing the formwork objects (productive actions) collected from a lexical database from English (WordNet)([27]) and no-interaction as non-productive actions. Then, to process the model and dataset, we use the high-performance hardware NVIDIA GeForce RTX 4090 with a GPU memory of 24.5GB and 131GB for RAM.

We trained our model in two steps: object detection and GNN for the HOI tasks. For the Faster R-CNN, we used some training parameters, such as 2500 epochs, batch size 20, and learning rate 0.001, and we achieved 0.83 for the mAP@0.5. Then, in the next step, we train our HOI model with these conditions (learning rates 1e-05, batch size 32, drop probability 0.3, and epochs 300. We evaluate the model with IoU more than 0.5, which means if the intersection over the union of each bounding box is more than 0.5, we consider it as the candidate for building the graphs. Moreover, we define the performance as a true positive if the three predicted HOI elements (humans, objects, and actions) are suitable with the ground truth, and we consider false positives if the model can predict the classes that are not available in our datasets. We achieve the 0.674, 0.556, and 0,632 mAP scores (Table 5) for the local area, global area, and average area simultaneously as the evaluation report for our datasets.

Figure 3. Installing formwork [16]

Table 2. Installing action prediction, Fig 3

Items	Humans	Objects	Interactions
1	H0 worker-1	00. rebar-local area	no interaction (0.001)
2	H0. worker-1	01. formwork-local area	installing (0.008)
3	H1. worker-2	02. formwork-local area	preparing (0.011)

Figure 4. Preparing formwork [16]

Table 3. Preparing action prediction, Fig 4

Items	Humans	Objects	Interactions
1	H0. worker-1	01. rebar-local area	no interaction (0.001)
2	H0. worker-1	02. formwork-local area	no interaction (0.011)
3	H0. worker-1	02. formwork-local area	preparing (0.493)

Figure 5. Transporting formwork[16]

From Figure 3 and Table 2. we can see that the workers interact with the nearest objects; for instance, H0 (worker 1) has an installing action with the formwork but no interaction with the steel rebar in the local area. Then, the worker (H1) interacts with the stacks of formwork as the

Table 4. Transporting action prediction, Fig 5

Items	Humans	Objects	Interactions
1	H0. worker-1	01. formwork-local area	transporting (0.015)
2	H0. worker-1	04. formwork-global area	preparing (0.010)
3	H1. worker-2	01. formwork-local area	preparing (0.010)
4	H1. worker-2	02. formwork-local area	transporting (0.05)
5	H1. worker-2	03. formwork-local area	preparing (0.266)

Table 5. Performance report of the HOI tasks

Actions	AP local area	AP global area	AP average area
Installing formwork	0.844	0.540	0.681
Preparing formwork	0.644	0.603	0.626
Transporting formwork	0.533	0.525	0.588
mAP score	0.674	0.556	0.632*

Table 6. mAP (%) score comparison with the previous studies in HICO-DET groups

HOI models	mAP average area
HO-RCNN [8]	0.078
PMN [28]	0.212
VS-Gats [10]	0.203
ViPLO [29]	0.372
PVIC-SwinL [30]	0.443
Ours	0.632

preparing action. In Figure 4 and Table 3, worker-1 (H0) mainly focuses on the formwork board or preparing the formwork structures, but the surrounding entities lack rebar rods and formwork components are also detected as the preparation, inspection, and no interaction activity. The last qualitative result is the transporting works (Figure 5, and Table 4); the two workers are collaborating in transporting the formwork's elements. Here, worker-1 (H0) transports the formwork-local area (01), while worker-2 (H1) transports the small sheet of board (formwork-local area 02) in a different area. In addition, we also compare the results with the existing HOI model as a validation process in which our average area (mAP 0.632) is higher 0.189 than PVIC-SwinL as the current highest performance HOI model (Table 6), especially in the HICO-DET domain. This result indicates that the model can comprehensively localize the objects on both large and small scales and predict the action connection accurately.

After we obtain and evaluate the results, managers can monitor the productivity of each worker. Productivity can be categorized by the type of detected action in the model. From the predicted results in each frame, we can conclude that they are productive workers with the installing, preparing, and transporting actions for the formwork structure as the major objects. Moreover, worker compliance, such as wearing standard safety attributes, using complete tools, and having a high dedication to working as professional workers, can also be considered by managers to determine the productivity of the workers. However, minor objects that do not have direct interaction with the workers, such as steel rebar or unused formwork, are predicted as no-interaction actions. This means the workers are not productive in minor objects but perform well in major objects. In addition, the prediction in the other frames can have different results; the major objects can be non-productive if the worker has no interaction actions at the prediction level. Therefore, to have an accurate analysis, Figure (6) shows the total number of the detected actions for all workers in the construction sites. Although the productive actions (preparing) are dominated in the frame, the prediction number of no-interaction actions is also high, so managers should arrange a better scenario to address this issue.

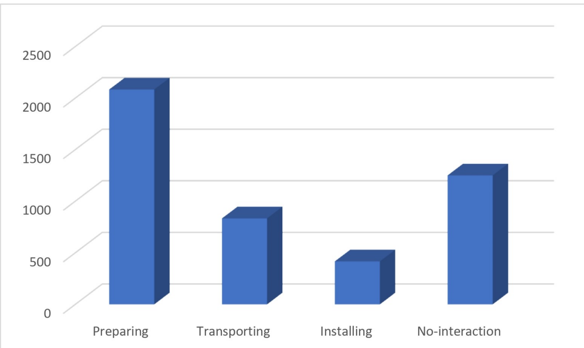

Figure 6. Productivity monitoring results

As the discussion part, even though we achieved excellent performance results, some limitations still exist. Our model has an imbalance data issue that prevents us from reaching better results. The public datasets we obtained do not represent the real conditions in construction sites. The current datasets lack variations, and our datasets might not perform well if a new set of data (unseen) is being used for the testing process. Moreover, our study focuses more on static images in which the dynamic interaction of the workers has not been addressed yet. To conclude, for future studies, we recommend utilizing more features, such as temporal features from the video or real-time scenario and body pose features, to improve the accuracy of the action prediction. Advanced methods, such as vision-language models (VLM), can also be utilized to improve the detection, apply the segmentation, solve the data imbalance, and process the semantic feature much better.

5 Conclusion

This study investigates the missing link or connection between the workers and objects that play an important role in measuring the productivity monitoring process. While the current studies focus solely on object detection and activity recognition only, we propose the HOI approach to address this issue. Our study achieves accurate results in both object detection (Step 1) and interaction prediction (Step 2). Thus, we can identify which workers are productive and non-productive using action recognition. However, the imbalanced data as the main prompt in this study must be addressed with current advanced technology and other effective scenarios. Thus, we can provide the level of the assessment process to improve the productivity level of the construction industry in the future.

Acknowledgment This research received support from four institutions: first, the Ministry of Education (MoE) of Taiwan provided full funding for the doctoral program. Second, the AI Research Center and the Department of Civil Engineering at National Taiwan University (NTU), contributed research funding and high-performance facilities. Third, the AI Center at Brawijaya University (UB) in Indonesia offered research partnership grants (95/UN10.F07.11/KS/2024) and provided annotation teams to assist in completing the study. In addition, the National Science and Technology Council (NSTC) of Taiwan (Case No. 114PGFA0100550) facilitated the international conference expenses. Thus, this study can be finished, accepted, and presented at ISARC 2025.

References

[1] Koen Vermeltfoort-Birgit Biemans Jan Mischke, Kevin Stokvis. Delivering on construction productivity is no longer optional. Technical report, McKinsey Company, 2024. URL `https://www.mckinsey.com/`.

[2] Khosiin, Lin, and Chen. Worker accountability in computer vision for construction productivity measurement: A systematic review.

Korea Institute of Construction Engineering and Management, pages 775–782, july 2024. doi:https://doi.org/10.6106/ICCEPM.2024.0775.

[3] Jiukun Dai, Paul M. Goodrum, William F. Maloney, and Cidambi Srinivasan. Latent structures of the factors affecting construction labor productivity. *Journal of Construction Engineering and Management*, 135(5):397–406, 2009. doi:10.1061/(ASCE)0733-9364(2009)135:5(397).

[4] O. Salem, J. Solomon, A. Genaidy, and I. Minkarah. Lean construction: From theory to implementation. *Journal of Management in Engineering*, 22(4):168–175, 2006. doi:10.1061/(ASCE)0742-597X(2006)22:4(168).

[5] Shang-Hsien Hsieh Mani Golparvar-Fard Aritra Pal, Jacob J. Lin. Activity-level construction progress monitoring through semantic segmentation of 3d-informed orthographic images. *Automation in Construction*, RR156-11:105157, 2024. ISSN 0926-5805. doi:https://doi.org/10.1016/j.autcon.2023.105157.

[6] Construction Industry Institute (CII). Engineering productivity measurement. *Construction Industry Institute*, RR156-11:286, 2001.

[7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions, 2018. URL `https://arxiv.org/abs/1704.07333`.

[8] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *CoRR*, abs/1702.05448, 2017. URL `http://arxiv.org/abs/1702.05448`.

[9] Hongsheng Li, Guangming Zhu, Wu Zhen, Lan Ni, Peiyi Shen, Liang Zhang, Ning Wang, and Cong Hua. Spatial parsing and dynamic temporal pooling networks for human-object interaction detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. doi:10.1109/IJCNN55064.2022.9892087.

[10] Zhijun Liang, Junfa Liu, Yisheng Guan, and Juan Rojas. Visual-semantic graph attention networks for human-object interaction detection. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1441–1447, 2021. doi:10.1109/ROBIO54168.2021.9739429.

[11] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi:10.1109/TNN.2008.2005605.

[12] Qilin Zhang, Zhichen Wang, Bin Yang, Ke Lei, Binghan Zhang, and Boda Liu. Reidentification-based automated matching for 3d localization of workers in construction sites. *Journal of Computing in Civil Engineering*, 35(6):04021019, 2021. doi:10.1061/(ASCE)CP.1943-5487.0000975.

[13] Ying Sun and Zhaolin Gu. Using computer vision to recognize construction material: A trustworthy dataset perspective. *Resources, Conservation and Recycling*, 183:106362, 2022. ISSN 0921-3449. doi:https://doi.org/10.1016/j.resconrec.2022.106362.

[14] Ronald Mutegeki and Dong Seog Han. A cnn-lstm approach to human activity recognition. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 362–366, 2020. doi:10.1109/ICAIIC48513.2020.9065078.

[15] Sagrario Garcia-Garcia and Raúl Pinto-Elías. Human activity recognition implenting the yolo models. In *2022 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*, pages 127–132, 2022. doi:10.1109/ICMEAE58636.2022.00029.

[16] Min-Yuan Cheng, Akhmad F.K. Khitam, and Hongky Haodiwidjaya Tanto. Construction worker productivity evaluation using action recognition for foreign labor training and education: A case study of taiwan. *Automation in Construction*, 150:104809, 2023. ISSN 0926-5805. doi:https://doi.org/10.1016/j.autcon.2023.104809.

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

[18] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu , changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek,

and Prashant Rai. ultralytics/yolov5: v3.1 - bug fixes and performance improvements, October 2020. URL https://doi.org/10.5281/zenodo.4154370.

[19] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 650–657, 2017. doi:10.1109/FG.2017.82.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. URL http://arxiv.org/abs/1703.06870.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735.

[22] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. *CoRR*, abs/1911.06644, 2019. URL http://arxiv.org/abs/1911.06644.

[23] Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, 15(6):455–469, 1982. ISSN 0031-3203. doi:https://doi.org/10.1016/0031-3203(82)90024-3.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[26] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[27] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11): 39–41, November 1995. ISSN 0001-0782. doi:10.1145/219717.219748.

[28] Zhijun Liang, Junfa Liu, Yisheng Guan, and Juan Rojas. Pose-based modular network for human-object interaction detection. *CoRR*, abs/2008.02042, 2020. URL https://arxiv.org/abs/2008.02042.

[29] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. ViPLO: Vision Transformer Based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17152–17162, Los Alamitos, CA, USA, 2023. IEEE Computer Society. doi:10.1109/CVPR52729.2023.01645.

[30] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10411–10421, October 2023.