Data-Driven Worker Detection from Load-View Crane Camera

Tanittha Sutjaritvorakul, Axel Vierling and Karsten Berns

Department of Computer Science, Technische Universität Kaiserslautern, Germany
E-mail: {tanittha,vierling,berns}@cs.uni-kl.de

Abstract -

Cranes as an essential part of the construction machinery, are one of the prominent sources of fatalities in the construction sites. The camera assistant system can contribute significantly to the safety of the crane operation particularly in blind lifts tasks, where the operator highly relies on the load-view camera. In this paper, we address the worker detection from an off-the-shelf load-view crane camera using a data-driven approach. Due to the difficulties in collecting data, we generate five training datasets via a simulation platform to build up the synthetic samples to improve the state-of-the-art detector. Despite the fact that only the simulation data is used as training datasets, the trained network demonstrates the average precision of up to 66.84% in two real-world scenarios.

Keywords -

Construction safety; Crane simulation; Worker detection; Visibility assistance

1 Introduction

The number of crane accidents caused by visibility remains high. The load-view crane camera is essentially used to widen the operator's field of v iew. However, it is hard for operators to keep observing hazards from merely a seven-inch monitor with no semantic information such as the position of the worker with respect to the crane or load. This work presents an analysis of a data-driven worker detection from a load-view camera using solely synthetic data in the learning procedure. The large volume of synthetic data is created by the simulation platform.

According to visibility-related fatalities, struck-by accidents contribute to 87.7% of all construction equipment accidents [8]. Cranes, which are the main machines in the construction, carry out the major activities in the building construction industry. Falling loads or struck-by loads are the most common and most dangerous crane-related hazards. The workers can be struck or hit by any moving load while they are working in close proximity to the crane. The poor visibility or blindspot causes the operator has difficulties to identify any personnel or objects in the work zone. Unlike in the street environment, the construction site is complex and unstructured. Workers and machines work side by side. The operator simultaneously monitors many things e.g., load radius, workers-on-foot, and spotter. Automated localizing workers or objects surrounding the load allows operators to understand the situation, and make decisions and actions accordingly.

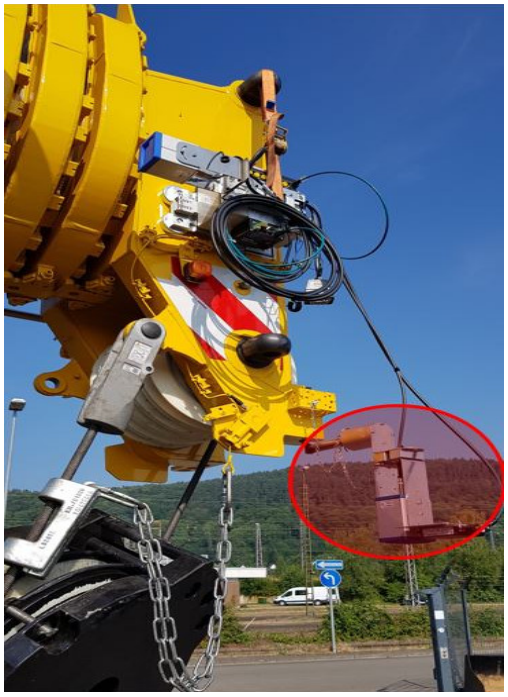

Figure 1. Crane load-view camera circled in red.

Not all sensor types are suitable to monitor objects from the load. Numerous crane safety assistance methods are presented in previous studies. The proximity warning is prevalent. Many sensor-based technologies have been adopted for construction safety assistance. These sensors can be installed on the site, workers or machines themselves to recognize objects. To increase spatial awareness of the operator, the load sway monitor can be observed using IMU or a camera [4, 5]. Similarly, hook motion tracking can measure the working radius in order to avoid collision [28, 12, 17]. Wearable devices such as bluetooth, RFID, and ultra-wideband (UWB) tag on safety helmet which provides the position can be irritating and privacy-intrusive to the workers [25, 29]. The operator mainly depends on the load-view camera during lifting tasks. It allows the operator to inspect the distance between the load and obstruction without occlusion among other objects. The view provides the top perspective from the camera mounted at the boom head pointing down to the ground. Information from any sensors installed on-site or on the cabin itself can be insufficient for the operator

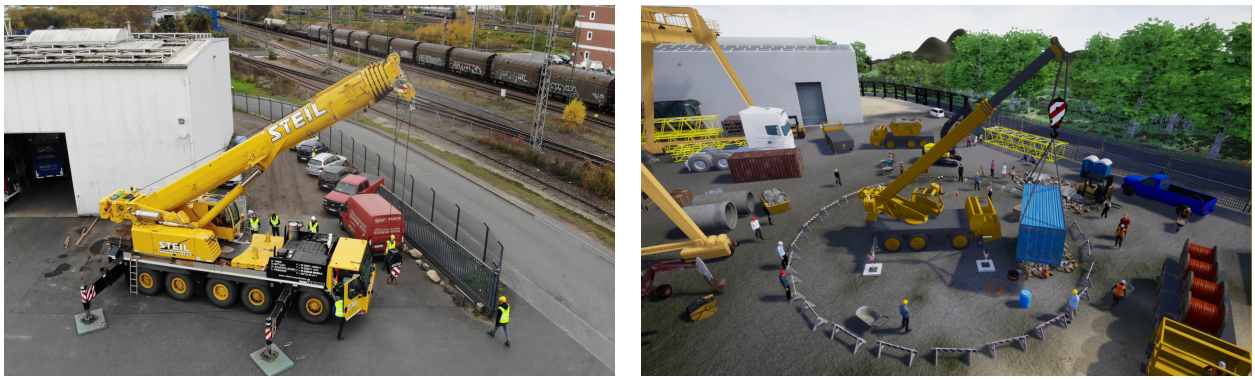

Figure 2. Comparison of real (left) and virtual (right) view at the experimental site in Trier, Germany.

during lifting materials over or into the building. The operator is unable to directly observe any adjacent objects due to the obstruction.

Detecting workers from the load-view is challenging. There is a lack of research on this topic. Traditional worker detection methods are based on simple features like helmet and color of high visibility vest [18, 26, 24]. In fact, most workers do not regularly wear protective clothes. The average of 87 percent of construction workers is reported as Personal Protective Equipment (PPE) noncompliance [20]. Thus using PPE information, the high-visibility color of helmet or vest, as a feature to detect workers may not be adequate. In general, it is very hard for a human to notice the small-sized workers from the top view, see Figure 3(f) which explains the high fatality rate in crane operation and necessitates the application of load-view worker detection. The construction area is cluttered and dynamically changing over time.

With the outstanding results, the construction domain also employs the data-hungry learning methods into worker recognition. The following studies of detecting workers from load-view camera or similar use deep learning approaches. Yang et al. [34] use Mask R-CNN to detect workers from a tower crane and identify if the workers are in the safety distance. Hu et al. [9] use YOLOv3 to detect non-complaint worker without a helmet. Vierling et al. [30] propose an automatic zoom load-view camera based on the working zone and load occlusion. The authors train the convolutional neural network with the load-view image and current zoom level, then result the optimal zoom level for the operator.

There is an intensive shortage of training data in the construction domain. The performance of deep learning methods is highly dependent on the existence of ample training samples. The self-driving car datasets publicly exist in great amount [11]. However, these datasets are not applicable to load-view object detection due to the frontal viewpoint. In addition, Unmanned Aerial Vehicle (UAV) datasets [23, 33, 1] could not be used as an alternative because of an uncluttered and static background, unlike the construction area. The pose or activity of the worker and the pedestrian are not identical, which can lead to different image features.

Data collection is crucial. Gathering data consists of two main steps, namely data recording, and annotation. Recording data from a car is relatively straightforward as opposed to a huge construction machine. The sensors can be easily mounted and adjusted. The driver does not require any additional specific skills. Image annotation techniques can be manual, semi-automatic, and automatic. Manual annotating data is tedious. The annotator required the knowledge to fulfill the task e.g., occlusion constraints, object representation, and boundary [2]. For a very large scale dataset, there exist crowd-sourcing platforms, such as Amazon Mechanical Turk (MTurk), to gather image annotation possible. Regardless of the verified annotated data, Zhang et al. [35] show the localization errors of original annotation in Caltech dataset.

Besides the benefits of using simulation as a construction robot test platform or vocational training, simulation also helps to augment data while reducing localization error and time from the manual labeling. Vierling et al. [31] develop the automated data generation tool in a game engine. Soltani et al. [27] propose automated annotation using synthetic images of construction resources is able to reduce the annotating time while improving the detection accuracy. The synthetic data can be used as an additional option to generate the training samples. Several studies [19, 32] demonstrate the synthetic data, which is generated from a game engine, can be used in a real-life scenario. With the rendering capability, the game engines like Unreal Engine[1] can generate the photorealistic environment and human characters. The virtual characters

[1]A game engine developed by Epic Games (www.unrealengine.com)

should behave naturally. Jan et al. [7] modelled and validated the usage of virtual characters in Unreal Engine for pedestrian-vehicle interaction system for an autonomous vehicle.

This paper aims to detect workers from a load-view crane camera using a data-driven detection approach. Worker detector can semantically provide information about what is happening nearby the load for the operators. With promising performance of DNN, RetinaNet architecture [14] is selected as a worker detector for our experiment. In order to cope with the absence of data, we generated the synthetic training data from virtual environment which is similar to the real experimental site. Special attention is given to construct the worker appearance, clothing and movements.

2 Methodology

Our approach consists of two main parts, data collection, and worker detection. First, the data collection describes how we gather the dataset from the real scenario and simulated platform. The second part explains the choice of detection algorithm and training strategy. The test crane used in this work is a telescopic crane (Liebherr LTM1130) while the testing took place in Trier, Germany. The standard crane load-view camera (Motec MC5200) is used in detecting workers. It is mounted at the boom end via pendulum bracket, looking downward, see Figure 1. The hardware used in detection experiments is NVIDIA GeForce GTX 1060, 3GB GDDR5.

(a) UT0 (b) UT1 (c) UT2 (d) UT4 (e) R1 (f) R2

Figure 3. Sample datasets for worker detection.

2.1 Data Collection

All collected data is listed in Table 1. The sequence name with a prefix of *R* is collected from the real mobile crane at the experimental site while the one with prefix *UT* is data generated by Unreal Engine. Examples of the data can be found in Figure 3. In a real-world scenario, the data is taken from the crane using 3-7 participants in the scene. The estimated distance of the camera to the ground (D_{cam}) is 25-35 meters, which refers to a 6- to 8-floor building. The crane performs basic lifting task—hoisting, extending, retracting boom, etc. In sequence R2, the test load is a wooden pallet. The annotation is done by hand which took about 14-20 seconds per frame.

For the synthetic data, we use a simulation system that developed in [32, 6]. The platform exploits the game engine features which allow us to create alike environment as the experimental site, see Figure 2. It provides large, diverse data with accurate annotation in an instant. The datasets contain workers, with and without PPE on diverse appearances and activities e.g., talking on the phone, standing upright, driving in the truck, carrying, pushing the wheelbarrow, or working with the device. Similar to the real world, construction machines, equipment, and material are present. Within the same scenario, different weather conditions can be rendered. The load-view camera setup is installed in the same manner as the real hardware. We generated 5 virtual image sequences, UT0-UT4 under daylight conditions. The sequence number of the synthetic data defines the level of boom arm extension e.g., UT0 means no boom extension and UT4 means the crane extends the boom up to 4th section. The main boom angle to the ground of all synthetic sequences is 60 degrees. In each sequence, the images are randomly captured while the turret is rotating from 0 to 360 degrees.

2.2 Worker Detection

Choosing network architecture is a difficult task because of speed–accuracy trade-offs [10]. With the great achievement of the deep neural network (DNN), it has been widely used and takes over the traditional image recognition methods. Regarding the requirement of visibility assistance, the operator should be alarmed about any surrounding

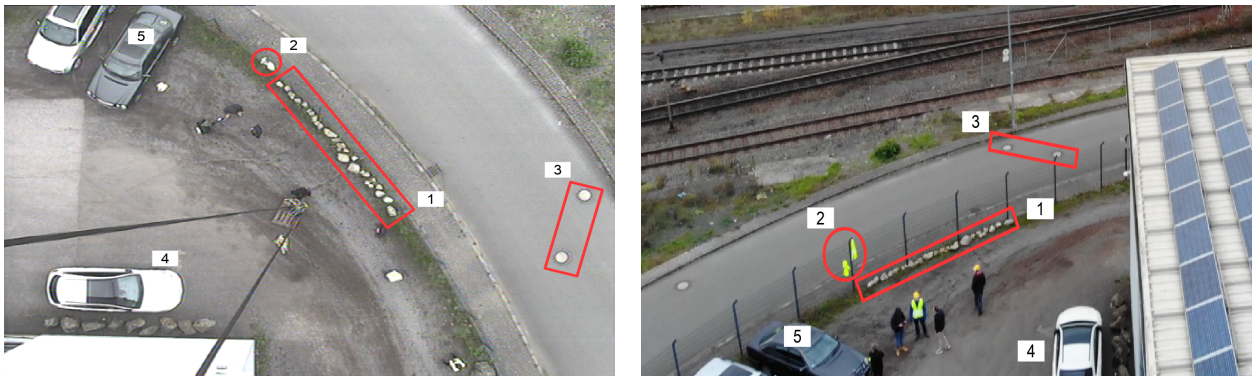

Figure 4. Comparison of the top view perspective between load-view camera (left) and drone camera (right). The identity of each object in both images is defined by the same number tag in the scenario. Number 1 is a rock border next to the fence. Number 2 is two green emergency vests hanging on the fence. Number 3 is two road manholes. Number 4-5 are cars.

Table 1. Dataset Summary.

Seq name	Frames	Resolution	D_{cam}(m)	Average object instances per frame	Total object instances
UT0	120	1600x1200	12	2	283
UT1	300	1600x1200	19	3	753
UT2	303	1600x1200	26	5	1636
UT3	501	1600x1200	33	9	4463
UT4	1110	1600x1200	39	8	8448
R1	713	720x480	25	3	2139
R2	400	720x480	35	7	2795

workers-on-foot in order to be aware of the hazards in (near) real-time.

Object detectors based on the DNN can be mainly categorized into two groups, two-stage detector, and single-stage detector. Two-stage detectors, such as all R-CNN model series [22], are mainly based on regional proposal network (RPN). In the first stage, the model proposes a set of sparse regions of interest by RPN or selective search. The candidates are later classified in the second stage. The accuracy of these models results relatively high but is typically slower. On the other hand, one-stage detectors, SSD [16], YOLO family models [21], and RetinaNet [14], propose the candidates from the input image directly without region proposal step. This leads to simpler and faster model architecture while lessening the performance slightly.

In this paper, we select the object detector based on RetinaNet for our experiments. It is introduced to handle objects in different scales and accurately localize dense objects. The focal loss in RetinaNet tackles an extreme imbalance between background which contains no object and foreground which has objects of interest. In other words, there are a very large number of negative samples and only a few positive samples. Therefore, RetinaNet works well in detecting small targets and high density such as the view from the UAV or load-view crane camera. The backbone network of RetinaNet is the featurized image pyramids which allow detecting object in multiscale [13].

To create the synthetic data closely resembling the target dataset (i.e., R1 and R2), the synthetic data are pre-processed by image filtering. We notice that the target images have more motion blur than the training samples because they tend to come from the swing movement of the camera, the vibration of the machine, or video interlace. For this reason, the motion blur is added to the synthetic data. In practice, the averaging filter with the kernel size of 10x10 applies to all simulated data in order to blur the images. The original synthetic datasets are denoted as UT0-UT4 and the blurred datasets are denoted $\overline{\text{UT0}}$-$\overline{\text{UT4}}$.

The ResNet-50 model is used as a backbone network. We initialize our weights from a pre-trained checkpoint of the COCO dataset [15]. All synthetic data, $\overline{\text{UT0}}$-$\overline{\text{UT4}}$ are combined and randomly shuffled into training and validation sets. The train set and the validation set consist of

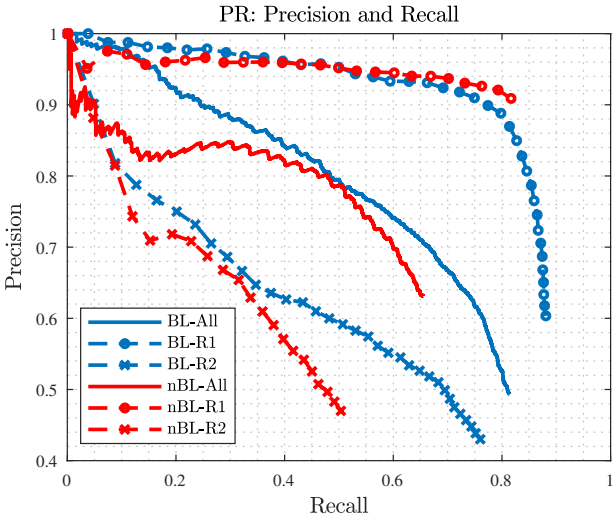

Figure 5. Precision-Recall curves of the experiments. AP in Table 2 can be achieved by the approximation of area under PR curve.

10907 and 4675 objects respectively. The test set with 4934 objects are from R1 and R2. The network is trained until the optimal point with a learning rate of $1e-7$. The sizes of anchors are set to {32, 64, 128, 256, 512} and the strides to {8, 16, 32, 64, 128}.

3 Results and Analysis

We conducted two main trials. In the first trial (BL), we trained the network with the blurred images ($\overline{\text{UT0}}$-$\overline{\text{UT4}}$) and for the second trial (nBL) the non-blurred images, UT0-UT4, are used for training. In each trial, we validate the network with two test sets, R1 and R2. Our detection evaluation metric is adopted from PASCAL Challenge [3] with Intersection over Union (IoU) threshold of 0.5.

$$IoU = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \tag{1}$$

B_p is predicted bounding box and B_{gt} is ground truth bounding box. Average Precision (AP) is widely used in measuring the accuracy among object detectors. The metric is based on the precision-recall (PR) curve. Figure 6 presents several predicted frames from both test sets. APs of the trials are listed in Table 2. The AP is obtained by the approximation of areas under PR curve.

First, we evaluate the networks, which are trained with blurred and non-blurred images on the test sequence R1. Both of them, $BL-R1$ and $nBL-R1$, yield nearly the same results (AP≈78%). The workers in the sequence most often can be recognized by both networks. Despite the low-light condition, the workers were wearing the high-visibility color vest and hard helmet which can be visible

Table 2. Results of AP on each dataset.

Trial	Test seq name	AP@0.5 (%)	Average inference time (ms per frame)
$BL-All$	R1,R2	**66.84**	-
$BL-R1$	R1	78.10	150.0
$BL-R2$	R2	**50.10**	152.7
$nBL-All$	R1,R2	53.13	-
$nBL-R1$	R1	**78.20**	155.6
$nBL-R2$	R2	38.26	151.7

to the networks. Afterward, we assessed the second test sequence R2 for the trial $BL-R2$ and $nBL-R2$. The detector trained with blurred images, $BL-R2$, shows a positive outcome. As a result, the overall AP of the network is higher when trained with the blurred datasets ($\overline{\text{UT0}}$-$\overline{\text{UT4}}$), compared to the non-blurred ones (UT0-UT4), check the AP values for trial $BL-All$ and $nBL-All$ in Table 2. The difference in the average predicting times among trails is minor.

In fact, R2 is a difficult sequence. It is recorded in higher elevation and thus it is hard to recognize the worker. Figure 4 shows the comparison of the same objects from two different camera angles. Apparently, the white rocks (number 1) and manholes (number 3) are almost identical to the person wearing the safety helmet. The workers' appearance form a similar color and shape view as of the ground. For the green emergency vest, we notice that the load-view camera is unable to reproduce the same color as shown in the drone camera or being visible to the human eye. Instead, it displays as white pixels, see Figure 4. This could be caused by the variant brightness, low image resolution, etc. In addition to the issue of the traditional detectors using only PPE color features mentioned in Section 1, color inaccuracy shown in the load-view camera can worsen these detectors because those color feature ranges are normally predefined. These negative samples can likely lure the human to misjudge as well as the detector.

Furthermore, we had prior experience in training the load-view worker detector with UAV data whose detail is not included in this work. The data are initially expected to be used as an alternative to augment the training dataset for load-view worker detection. However, the prediction results were unsatisfactory. Evidently, the workers in the drone camera in Figure 4 can be seen fully while only the heads and shoulders of the workers in the load-view camera are visible. Consequently, using artificial data to train a DNN model is beneficial. The model acquires the image features and is able to yield good performance without seeing none of the real-world data.

Figure 6. Predicted results of trial *BL* on the test sequence R1 in the first row (frame 30, 219, 632) and R2 in the second row (frame 40, 297, 348). The blue bounding box is the detected target with confidence score label while the green box is groundtruth.

4 Conclusion

In this paper, we demonstrate the worker detector from a load-view camera using RetinaNet. This one-stage detector is able to localize and classify small-sized objects in dense areas. Two test image sequences are collected from the real crane. Regarding data shortage and complexity in data collection, we created the five image sequences from different altitudes by the simulated platform in Unreal Engine. The platform allows us to generate plenty of data in an accurate and fast manner. The datasets are synthesized with the motion blur and later fed into the learning procedure. There are two networks trained for evaluation. The first network is trained with preprocessed images and the second is trained with the primitive images. Finally, the detector ran on the two test sequences were taken from the real crane. Blurred virtual data appears to make data more realistic to the learning algorithm.

For future study, worker tracking and activity recognition could be added to reduce misinterpretation between non-object and object. Different synthesized techniques can possibly experiment on the images for training, such as video interlace and synthetic image refiner. Using synthetic data still requires more effort to study because the synthetic data sometimes looks realistic to a person but it can appear to be unrealistic to the learning algorithms.

In conclusion, the worker detector can be used as additional information for risk assessment for each worker. Visualization of workers nearby in 2D or 3D with respect to the crane including the risk level of each worker can be useful for the situational awareness of the operators. This can provide support to the crane operators to identify hazards during operation.

5 Acknowledgement

This work is funded from the Federal Ministry of Education and Research (BMBF) under grant agreement number 01|16SV7738 and named SAFEGUARD.

References

[1] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 28–35, 2017.

[2] Adela Barriuso and Antonio Torralba. Notes on image annotation. arXiv preprint arXiv:1210.3448, 2012.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[4] Yihai Fang, Jingdao Chen, Yong K Cho, Kinam Kim, Sijie Zhang, and Esau Perez. Vision-based load sway monitoring to improve crane safety in blind lifts. Journal of Structural Integrity and Maintenance, 3(4):233–242, 2018.

[5] Yihai Fang and Yong K Cho. Crane load positioning and sway monitoring using an inertial measurement unit. In Computing in Civil Engineering 2015, pages 700–707. 2015.

[6] Jan Qazi Hamza, Kleen Jan, and Karsten Berns. Self-aware pedestrians modeling for testing autonomous vehicles in simulation. In Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2020), 2020.

[7] Jan Qazi Hamza, Klein Sascha, and Berns Karsten. Safe and efficient navigation of an autonomous shuttle in a pedestrian zone. In International Conference on Robotics in Alpe-Adria Danube Region, pages 267–274. Springer, 2019.

[8] Jimmie W Hinze and Jochen Teizer. Visibility-related fatalities related to construction equipment. Safety science, 49(5):709–718, 2011.

[9] J. Hu, X. Gao, H. Wu, and S. Gao. Detection of workers without the helments in videos based on yolo v3. In 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 1–4, 2019.

[10] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7310–7311, 2017.

[11] Charles-Éric Noël Laflamme, François Pomerleau, and Philippe Giguère. Driving datasets literature review. arXiv preprint arXiv:1910.11968, 2019.

[12] Yanming Li, Shuangyuan Wang, and Bingchu Li. Improved visual hook capturing and tracking for precision hoisting of tower crane. Advances in Mechanical Engineering, 5:426810, 2013.

[13] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.

[17] Shunsuke Nara, Daisuke Miyamoto, and Satoru Takahashi. Position measurement of crane hook by vision and laser. In IECON 2006-32nd Annual Conference on IEEE Industrial Electronics, pages 184–189. IEEE, 2006.

[18] M Neuhausen, J Teizer, and M König. Construction worker detection and tracking in bird's-eye view camera images. In Proceedings of the 35th ISARC, Berlin, Germany, 2018.

[19] Marcel Neuhausen, Patrick Herbers, and Markus König. Synthetic data for evaluating the visual tracking of construction workers. In Construction Research Congress 2020, 2020.

[20] Occupational Health and Safety. Survey Finds High Rate of PPE Non–Compliance —Occupational Health and Safety, November 2008. [Online; accessed 18-06-2019].

[21] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. CoRR, abs/1804.02767, 2018.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 91–99. Curran Associates, Inc., 2015.

[23] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In European conference on computer vision, pages 549–565. Springer, 2016. Stanford Drone Dataset.

[24] A. H. M. Rubaiyat, T. T. Toma, M. Kalantari-Khandani, S. A. Rahman, L. Chen, Y. Ye, and C. S. Pan. Automatic detection of helmet uses for construction safety. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), pages 135–142, 2016.

[25] Suranga Seneviratne, Yining Hu, Tham Nguyen, Guohao Lan, Sara Khalifa, Kanchana Thilakarathna, Mahbub Hassan, and Aruna Seneviratne. A survey of wearable devices and challenges. IEEE Communications Surveys & Tutorials, 19(4):2573–2620, 2017.

[26] H Seong, H Choi, H Cho, S Lee, H Son, and C Kim. Vision-based safety vest detection in a construction scene. In ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, volume 34. IAARC Publications, 2017.

[27] Mohammad Mostafa Soltani, Zhenhua Zhu, and Amin Hammad. Automated annotation for visual recognition of construction resources using synthetic images. Automation in Construction, 62:14–23, 2016.

[28] Satoru Takahashi and Shun'ichi Kaneko. Motion tracking of crane hook based on optical flow and orientation code matching. In 2008 10th IEEE International Workshop on Advanced Motion Control, pages 149–152. IEEE, 2008.

[29] D Triantafyllou, S Krinidis, D Ioannidis, IN Metaxa, C Ziazios, and D Tzovaras. A real-time fall detection system for maintenance activities in indoor environments. IFAC-PapersOnLine, 49(28):286–290, 2016.

[30] A Vierling, T Sutjaritvorakul, and K Berns. Crane safety system with monocular and controlled zoom cameras. In ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, volume 35, pages 1–7. IAARC Publications, 2018.

[31] Axel Vierling, Tanittha Sutjaritvorakul, and Karsten Berns. Dataset generation using a simulated world. In International Conference on Robotics in Alpe-Adria Danube Region, pages 505–513. Springer, 2019.

[32] Sutjaritvorakul T. Vierling A., Pawlak J. and Berns K. Simulation platform for crane visibility safety assistance. In Proceedings of the 29th Conference on Robotics in Alpe-Adria-Danube Region (RAAD 2020), volume 84 of Mechanisms and Machine Science. Springer, Cham, 2020.

[33] Dongfang Yang, Linhui Li, Keith Redmill, and Ümit Özgüner. Top-view trajectories: A pedestrian dataset of vehicle-crowd interaction from controlled experiments and crowded campus. arXiv preprint arXiv:1902.00487, 2019.

[34] Zhen Yang, Yongbo Yuan, Mingyuan Zhang, Xuefeng Zhao, Yang Zhang, and Boquan Tian. Safety distance identification for crane drivers based on mask r-cnn. Sensors, 19(12):2789, 2019.

[35] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1259–1267, 2016.