Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

CONSTRUCTION IMAGE SYNTHETIZATION TO OVERCOME A SMALL, BIASED REAL TRAINING DATASET FOR DNN-POWERED VISUAL SCENE UNDERSTANDING

Jinwoo Kim

School of Civil and Environmental Engineering
Nanyang Technological University
50 Nanyang Avenue
Singapore, SG 639798, SINGAPORE

Daeho Kim

Department of Civil and Mineral Engineering
University of Toronto
35 St. George Street
Toronto, ON M5S 1A4, CANADA

SangHyun Lee

Department of Civil and Environmental Engineering
University of Michigan
2350 Hayward Street
Ann Arbor, MI 48109, USA

ABSTRACT

Deep neural networks (DNNs) have become a driving factor of visual scene understanding. However, the shortage of construction training images has been a major barrier to fully leverage its maximum performance potential. To address this issue, we investigate the effectiveness of synthetic images on DNN training in a common real-world scenario where only a small, biased real training image dataset is available. To this end, we synthetize numerous construction training images and conduct a DNN training experiment in real construction settings. Results show that the combined dataset-trained model always outperforms the one trained with only a small, biased real dataset. This finding indicates that an image synthetization approach has promising potential to enhance a given real training dataset in terms of data quantity and diversity. Image synthetization with automated labeling will mitigate the training image shortage, contributing to the development of more accurate and scalable DNNs for construction scene understanding.

1 INTRODUCTION

Robotic automation and digitalization has emerged as an effective means of improving construction productivity and safety (Kim et al. 2020a; Kim and Chi 2021). In the near future, robotic systems will be responsible for physically demanding, laborious, and dangerous operations (Kim et al. 2022), freeing human workers to focus more on supervising robots or troubleshooting uncertain events (Liang et al. 2021). Mobile robots that can digitally record, store, and analyze as-is onsite information will also help to achieve the digitalization of ongoing field operations (Davila Delgado et al. 2019). These capabilities will enable continuous, accurate, and digital inspection and control of field operations, leading to improved productivity and safety. These benefits of robotic automation will reform existing labor-intensive construction into a more innovative industry.

Visual scene understanding of complex and ever-changing workplaces is a central part of robotic automation and digitalization (Kim et al. 2019; Kim 2020). This will enable autonomous robots to place

themselves at as-planned working positions, perceive surrounding objects to reach or bypass, and adapt to changes in construction workplaces over time (Liang et al. 2021). However, as deep neural networks (DNN) have become a crucial basis of visual scene understanding, it is necessary to assemble an extensive construction training image dataset with accurate ground truth labels (e.g., object locations within an image) (Hwang et al. 2022; Kim et al. 2020b). Since manual training dataset assembly is time-consuming, laborious, and expensive, DNN training images used in construction studies have been limited in terms of data quantity and diversity (Kim and Chi 2022). This data shortage has held DNN models back from reaching their maximum performance potential for construction scene understanding (Kim et al. 2021).

To address this issue, we present an image synthetization approach that can automatically manipulate a wide spectrum of construction scenes (e.g., object poses, illuminations, and camera viewpoints) and create new artificial training images in a virtual environment. Being able to synthesize and label limitless construction training images without site visits and human effort will be innovative. To this end, we assemble a synthetic construction image dataset using Synthetic Human for Real Tasks (SURREAL) (Varol et al. 2017) and conduct a DNN training experiment, answering the following question: Can synthetic construction images offset the shortage of real training images? Specifically, we consider a common real-world scenario wherein only a small, biased real training image dataset is available. This study particularly targets construction worker detection since it is an essential task for construction scene understanding and this capability enables us to obtain workers' presence and location who are the key players in both current and future co-robotic construction.

2 EXISTING CONSTRUCTION BENCHMARK DATASETS AND DNN APPROACHES

Computer vision studies have made large strides in DNN-powered visual scene understanding, especially when it comes to scalable, comprehensive benchmark image datasets. Such successful achievements and technological advances have motivated construction researchers to develop and share construction-specific benchmark image datasets. (Kim et al. 2018), for example, extracted 2,920 images of heavy equipment from ImageNet, a well-known and popular benchmark dataset in the computer vision community, and reformed them as a construction-specific image dataset. (Xiao and Kang 2021) also released a construction dataset composed of 10,000 images corresponding to 10 different types of equipment (e.g., excavators, mixer trucks, and tower cranes). Another research group recently publicized a more comprehensive benchmark dataset called Moving Objects in Construction Sites (MOCS), consisting of 41,668 onsite images of 13 different construction objects (Xuehui et al. 2021). Although several studies have endeavored to build benchmark datasets for construction scene understanding, DNN model development in the construction domain falls behind the computer vision community due to the large gap between the quantity and diversity of training image datasets available in each field. For instance, while computer vision researchers have actively published a variety of multipurpose, large-scale benchmark datasets (e.g., object detection and segmentation, pose estimation, action recognition, depth estimation, and object-to-object relationship detection), there are only few available datasets for limited purposes in the construction domain (e.g., object detection and segmentation) (Xiao and Kang 2021; Xuehui et al. 2021). Moreover, such construction training image datasets, developed at a few specific jobsites, may have limited applicability in new construction environments with their own unique visual characteristics and imaging conditions. It is also forbidden to release construction site images in many cases because they can involve confidential and privacy information. This shortage of construction training images has: (i) resulted in DNN models that misbehave and malfunction in real-world construction sites; (ii) missed the chance to explore deeper and wider DNN architectures; and (iii) blocked finding out the optimal trade-off between the size of training datasets and the complexity of DNN architectures.

There have been a few studies to address the shortage of construction training images. Of these, virtual image synthetization has especially garnered increased attention as this process can automatically simulate diverse scene contexts (e.g., object poses and camera viewpoints) and create numerous synthetic images. (Braun and Borrmann 2019) synthesized virtual images of building elements (e.g., columns, walls, and slabs) using building information modeling. Their synthesized images were then used to train a vision-

based building element detection model and the model worked well in real-world construction scenarios. Similarly, a 3D virtual equipment model was leveraged to synthesize construction training images. (Assadzadeh et al. 2022) adopted a 3D virtual excavator model and simulated its physical behaviors to develop a synthetic dataset for 2D excavator pose estimation. This method was made applicable to 3D excavator pose estimation (Mahmood et al. 2022) and activity recognition in other studies (Torres Calderon et al. 2021). Despite their promising results, there has been a major roadblock to fully leveraging the benefits of synthetic images for DNN training due to the following knowledge gaps. First, it is questionable whether synthetic images can offset the data shortage when there are only a small number of real construction training images with biased scene contexts. Although this is a common real-world scenario due to limited data collectivity and accessibility in practice, existing studies have not yet investigated how synthetic images can supplement and strengthen the limited quantity and diversity of a small, biased real image dataset. Second, the effects of the number of synthetic images on DNN performance is still unknown in the aforementioned scenario. It is recognized that the major advantage of image synthetization is to create and label limitless virtual images, but earlier works focused on training DNN models with a fixed number of training input. Last, since previous studies focused mainly on building elements and heavy equipment, the positive effects of synthetic images with human workers remain unclear.

To fill these knowledge gaps, we aim to thoroughly investigate the effectiveness of synthetic images on DNN training, especially when a small, biased real training dataset is given. To this end, we (i) adopt and refine an image synthetization approach, i.e., SURREAL, and build a synthetic construction image dataset; (ii) combine a given small, biased real training dataset with the synthesized images; and (iii) train a DNN model with the combined dataset and evaluate its performance in real construction settings.

3 CONSTRUCTION IMAGE SYNTHETIZATION

In this phase of our research, we adopted and refined the SURREAL image synthetization framework (Varol et al. 2017) to assemble a virtual construction image dataset mainly capturing construction workers. Figure 1 illustrates the two main processes behind this phase: (i) construction worker modeling; and (ii) 2D image generation and automated labeling. Technical details of each process are described in the following subsections.

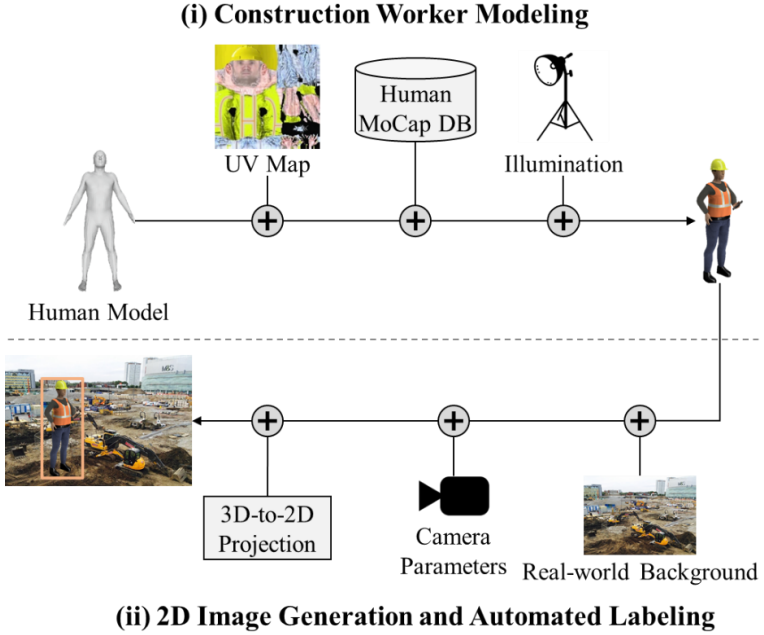

Figure 1. Overview of construction image synthetization framework.

3.1 3D virtual worker simulation

In this process, we designed a 3D virtual construction worker model and animated their kinematic behaviors in a virtual computer environment via four main steps. First, we adapted a well-known and precise human body shape model: a skinned multi-person linear (SMPL) model (Loper et al. 2015). The SMPL is a realistic human body model developed by 1,786 high-quality 3D scan data of real-world human subjects with a variety of Body Mass Index, including both males and females. The validity of SMPL has been proven for various visual scene understanding tasks, such as human detection and pose estimation (Bogo et al. 2016). Next, we refined the SMPL to be a more construction worker-looking model because it was originally built for humans in their daily lives (e.g., building occupants). In detail, we textured human skin by sketching a UV map for construction worker clothing, which is a 2D matrix that involves skin-texture information about the vertex of a 3D object model. As shown in Figure 1, hardhats and safety vests were built in the UV map. Third, we animated a wide spectrum of worker behaviors to diversify the visual characteristics of our construction worker model. The Carnegie Mellon University Graphics Lab Motion Capture Database (CMU Graphics Lab 2008) was adopted, containing more than 2,000 video-streams of 23 different types of human behaviors (e.g., standing, bending, and walking). As there were more than ten hours of 3D human behavioral video, it was possible to generate a large amount of synthetic worker images. Last, we varied the illumination of a synthetic 3D worker model using Spherical Harmonics with nine variables. The variables were randomly selected from a uniform distribution-based probability function between -0.7 and 0.7. This setting enabled us to simulate different levels of illumination in outdoor construction workplaces.

3.2 2D image capturing and augmentation

Using the 3D worker model, this process automatically captured a set of 2D synthetic images and generated label information via three main tasks. First, we overlaid the one single 3D worker model onto real-world construction images as shown in Figure 1. The use of real-world images as the backgrounds of virtual environments enabled DNNs to learn how to localize target objects (e.g., workers) in real construction settings while distinguishing them from other non-target objects (e.g., surrounding environments). In this study, a total of 529 construction images (without real workers) were collected from a website (e.g., Google) and used as the backgrounds. Next, we installed a virtual camera and captured 2D worker images in the well-controlled environment. As cameras can be installed at any distance with any viewpoint in a virtual environment, we were able to easily capture a great number of training images with diverse construction worker contexts (e.g., sizes and viewpoints). Camera distances were set to be randomly selected from a Gaussian distribution function (μ=30 m, σ=1 m) and camera viewpoints were set to be randomly selected from a range between 0 and π. This setting enabled us to obtain training images with varying image-capture conditions although the camera's intrinsic parameters (e.g., focal length and principle points) were fixed. Last, the virtual worker model was automatically labeled in the captured images by applying a 3D-to-2D projection algorithm, which translates 3D virtual-world coordinates into 2D image pixel coordinates. Conceptually, known 3D coordinates $[X, Y, Z]$ can be geometrically transformed into 2D pixel coordinates $[x, y]$ using a virtual camera's intrinsic (i.e., focal lengths and principal points) and extrinsic parameters (i.e., rotation angles and translations), which are obtainable in a computer environment. As a result, a total of 20,386 construction worker images with the label information (i.e., bounding boxes) were synthesized.

4 EXPERIMENT, RESULTS, AND DISCUSSION

We conducted an experiment to investigate if synthetic construction images can offset the shortage of real training images. Specifically, we assumed a common real-world application scenario wherein only a small, biased real training image dataset was available. In this scenario, we targeted construction worker detection because human workers are the main players in both current construction projects and future co-robotic construction and because object detection is an essential visual understanding task for robotic automation and digitalization. It should be also noted that, however, our image synthetization approach can be easily extensible to other types of construction objects (e.g., excavators and dump trucks) and visual scene

understanding tasks (e.g., pose estimation and activity recognition). The third version of you only look once (YOLOv3) was selected as a DNN architecture since its validity has been proven by existing studies (Kim et al. 2019). The DNN's performance was evaluated with an average precision (AP_{worker}), which is one of the most commonly used metrics in object detection tasks.

In this setting, we conducted an experiment with three cases to generalize our results: when 500, 1,000, and 2,000 real training instances (i.e., bounding boxes for workers) are given. In all cases, the real training dataset was combined with our synthesized construction images. In more detail:

- Training dataset (Case #1: 500 real instances): Of the publicly available MOCS dataset, we randomly selected 500 real worker instances. To strengthen this small, biased real dataset, we randomly sampled 1,000–10,000 synthetic worker instances at 1,000 intervals from our synthetic dataset and added them as supplementary training images.
- Training dataset (Case #2: 1,000 real instances): All settings were identical to Case #1, except that the number of real worker instances was set to be 1,000.
- Training dataset (Case #3: 2,000 real instances): All settings are identical to Cases #1 and #2, except that the number of real worker instances was set to be 2,000.
- Test dataset (Cases #1, #2, and #3): We randomly chose images having 10,000 worker instances from the publicized MOCS dataset, excluding the images included in the real training datasets from Cases #1, #2, and #3. Note that this test dataset was used in all cases.
- Weight initialization: The initial weights of our YOLOv3 architecture were set to be the weights of the model pretrained by the Common Objects in Contexts Dataset (Lin et al. 2014).
- Training-related hyper-parameters: We adopted an Adam optimization algorithm with a batch size of 8 and an epoch of 50. The learning rate was determined with a cosine annealing learning rate scheduler with initial and end values of 1e-4 and 1e-6.

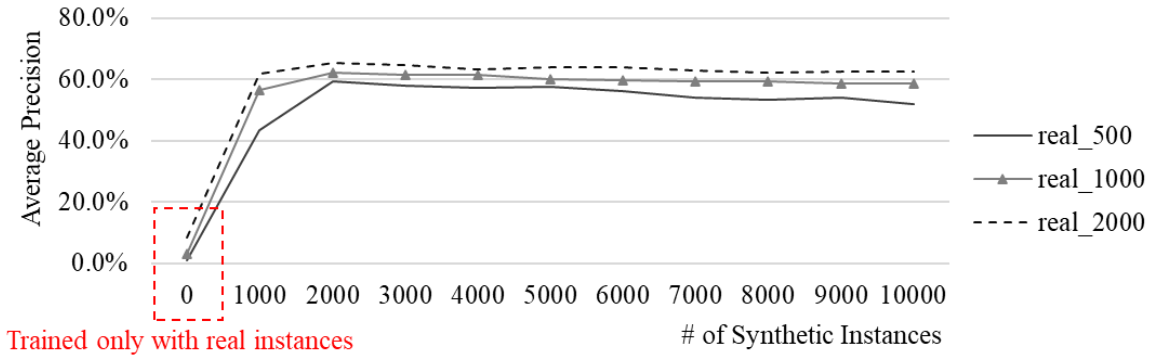

Figure 2. Model performance by the number of synthetic instances.

Figure 2 shows the model performance by the number of synthetic training instances for all three cases. Compared to when adopting only real images ($x = 0$ in the graph), the additional use of synthetic images could improve the DNN's performance from a minimum of 43.4% to a maximum of 59.0%. This remarkable improvement can be a clear indicator that synthetic images are able to supplement and strengthen a given small, biased real training dataset in terms of data quantity and diversity. In addition, it was observed that the maximum performance was achieved at 2,000 synthetic instances in all cases (real_500, real_1000, and real_2000) (Figure 2). It seems that 2,000 synthetic instances were sufficient to secure the quantity and diversity of a training image dataset.

Figure 3 illustrates that the model trained with 2,000 real and 2,000 synthetic instances performed well in outdoor and complex construction scenes. For example, the model could correctly localize multiple workers with varying poses and in cluttered backgrounds. The model also succeeded in detecting target workers under severe occlusions and in crowded scenes. These capabilities may be attributed to the fact

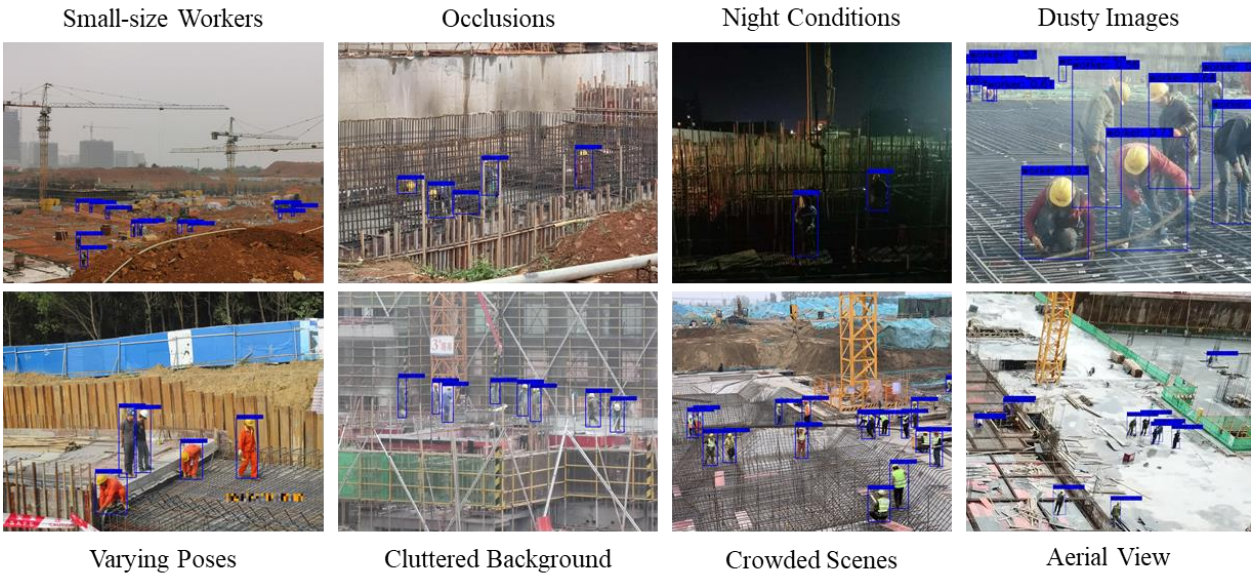

Figure 3. Examples of construction worker detection by the model trained with real and synthetic construction images.

that the synthetic 3D worker model and virtual environments were carefully controlled and the 2D images were captured at different locations and viewpoints. These results denote that, although there is only a small, biased real training dataset in practice, a promising DNN model can be developed by combining it with our synthetic images.

It was also interesting that the mixture use of real and synthetic training images always yielded a higher DNN performance than when using only real images. For example, the performance of the mixed image-trained model (500 real and 2,000 synthetic instances) was 59.5%, whereas the real image-trained one (2,000 real instances) remained in the vicinity of 8.3% (Figure 2). This result indicates that the supplementary use of synthetic images can significantly improve DNN performance by 51.2% while reducing the number of real instances by 75%, from 2,000 to 500. This finding suggests that practitioners could combine a small, biased real training dataset with synthetic images, rather than expending resources and time to gather and label real onsite images.

As shown in Figure 2, DNN performance peaked at 2,000 synthetic instances in all cases and then started to decline steadily when 3,000 or more synthetic instances were added to a given real training dataset. Particularly, the decrease in performance was more significant when (i) fewer real training instances and (ii) more synthetic instances were given. For instance, the performance dropped the most when combining 500 real instances with 10,000 synthetic ones (7.5% dropped). This result may be associated with adverse effects caused by the gap in visual characteristics between real and synthetic images. If the two different data sources with different characteristics are combined for training, a DNN model can become confused and disoriented. This finding implies that, when enlarging a small, biased real training dataset, we should carefully control the number of synthetic training images to reduce adverse effects on DNN performance.

5 CONCLUSION

We proposed an image synthetization approach and investigated the potential of synthetic images as DNN training images in a common real-world scenario: there is only a small, biased real training image dataset available. Specifically, we created synthetic construction images in a virtual computer environment; added the synthesized images to supplement and strengthen a small, biased real training dataset; and trained a DNN model with the combined dataset and evaluated its performance in real construction settings. In the

experiment, it was observed that the model trained with the combined dataset always had a much higher performance than the one trained with a small, biased real dataset. Quantitatively, performance improvement was reported with an average of 58.1%. This result shows that synthetic images have promising potential in enhancing a small, biased real training dataset in terms of data quantity and diversity. With these technical benefits, the combined dataset-trained model could successfully localize construction workers in diverse real-world settings (e.g., small-size workers, varying poses, night conditions, occlusions, and cluttered backgrounds) (Figure 3). This finding is more noteworthy since construction image synthetization and labeling can be fully automated in a virtual environment. Additionally, our approach can be flexible and applicable to many different types of visual scene understanding tasks, such as activity recognition, 2D/3D pose estimation, and depth estimation. These benefits can help build multipurpose, construction-specific training image datasets and ensure the accuracy and scalability of DNN models for visual scene understanding.

ACKNOWLEDGMENTS

We would like to thank Julianne Shah, who helped build the synthetic image dataset and the research group at Tsinghua University (ANLAB), who publicized the MOCS benchmark dataset. This research was supported financially by a National Science Foundation Award (No. IIS-1734266, 'Scene Understanding and Predictive Monitoring for Safe Human-Robot Collaboration in Unstructured and Dynamic Construction Environments').

REFERENCES

Assadzadeh, A., M. Arashpour, I. Brilakis, T. Ngo, and E. Konstantinou. 2022. "Vision-Based Excavator Pose Estimation Using Synthetically Generated Datasets with Domain Randomization". *Automation in Construction* 134:104089.

Bogo, F., A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. 2016. "Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image". *Lecture Notes in Computer Science* 9909:561–578.

Braun, A., and A. Borrmann. 2019. "Combining Inverse Photogrammetry and BIM for Automated Labeling of Construction Site Images for Machine Learning". *Automation in Construction* 106:102879.

CMU Graphics Lab. 2008. "CMU Graphics Lab Motion Capture Database". http://mocap.cs.cmu.edu/, accessed April 7th, 2021.

Davila Delgado, J. M., L. Oyedele, A. Ajayi, L. Akanbi, O. Akinade, M. Bilal, and H. Owolabi. 2019. "Robotics and Automated Systems in Construction: Understanding Industry-Specific Challenges for Adoption*". Journal of Building Engineering* 26: 100868.

Hwang, J., J. Kim, S. Chi, and J. Seo. 2022. "Development of Training Image DB Using Web Crawling for Construction Site Monitoring". *Automation in Construction* 135:104141.

Kim, D., S. Lee, and V. R. Kamat. 2020a. "Proximity Prediction of Mobile Objects to Prevent Contact-Driven Accidents in Co-Robotic Construction". *Journal of Computing in Civil Engineering* 34(4):04020022.

Kim, D., M. Liu, S. Lee, and V. R. Kamat. 2019. "Remote Proximity Monitoring between Mobile Construction Resources Using Camera-Mounted UAVs". *Automation in Construction* 99:168–182.

Kim, H., H. Kim, Y. W. Hong, and H. Byun. 2018. "Detecting Construction Equipment Using a Region-Based Fully Convolutional Network and Transfer Learning". *Journal of Computing in Civil Engineering* 32(2):04017082.

Kim, J. 2020. "Visual Analytics for Operation-Level Construction Monitoring and Documentation: State-of-the-Art Technologies, Research Challenges, and Future Directions". *Frontiers in Built Environment* 6:575738.

Kim, J., and S. Chi. 2021. "A Few-Shot Learning Approach for Database-Free Vision-Based Monitoring on Construction Sites". *Automation in Construction* 124:103566.

Kim, J., and S. Chi. 2022. "Graph Neural Network-Based Propagation Effects Modeling for Detecting Visual Relationships among Construction Resources". *Automation in Construction* 141:104443.

Kim, J., J. Hwang, S. Chi, and J. Seo. 2020b. "Towards Database-Free Vision-Based Monitoring on Construction Sites: A Deep Active Learning Approach". *Automation in Construction* 120:103376.

Kim, J., D. Kim, J. Shah, and S. Lee. 2021. "Training a Visual Scene Understanding Model Only with Synthetic Construction Images". In *Proceedings of 2021 International Conference on Computing in Civil Engineering*. September 12th-14th,

Orlando, Florida, 221-229.

Kim, Y., H. Kim, R. Murphy, S. Lee, and C. R. Ahn. 2022. "Delegation or Collaboration: Understanding Different Construction Stakeholders' Perceptions of Robotization". *Journal of Management in Engineering* 38(1):04021084.

Liang, C.-J., X. Wang, V. R. Kamat, and C. C. Menassa. 2021. "Human–Robot Collaboration in Construction: Classification and Research Trends". *Journal of Construction Engineering and Management* 147(10):03121006.

Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. "Microsoft COCO: Common Objects in Context". *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8693:740–755.

Loper, M., N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. 2015. "SMPL: A Skinned Multi-Person Linear Model". *ACM Transactions on Graphics* 34(6):1–16.

Mahmood, B., S. U. Han, and J. Seo. 2022. "Implementation Experiments on Convolutional Neural Network Training Using Synthetic Images for 3D Pose Estimation of an Excavator on Real Images". *Automation in Construction* 133:103996.

Torres Calderon, W., D. Roberts, and M. Golparvar-Fard. 2021. "Synthesizing Pose Sequences from 3D Assets for Vision-Based Activity Analysis". *Journal of Computing in Civil Engineering* 35(1):04020052.

Varol, G., J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. 2017. "Learning from Synthetic Humans". In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. July 21st-26th, Honolulu, HI, USA, 109-117.

Xiao, B., and S.-C. Kang. 2021. "Development of an Image Data Set of Construction Machines for Deep Learning Object Detection". *Journal of Computing in Civil Engineering* 35(2):05020005.

Xuehui, A., Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei. 2021. "Dataset and Benchmark for Detecting Moving Objects in Construction Sites". *Automation in Construction* 122:103482.

AUTHOR BIOGRAPHIES

JINWOO KIM is an assistant professor at Nanyang Technological University, Singapore. He earned his Ph.D. in the Department of Civil and Environmental Engineering at Seoul National University, Korea. His research interests center on construction digitalization and project performance analytics via AI-assisted data-driven approaches. His e-mail address is jinwoo.kim@ntu.edu.sg. His website is https://sites.google.com/view/jinwoo-kim/.

DAEHO KIM is an assistant professor at the University of Toronto, Toronto, Canada. He earned his Ph.D. in the Department of Civil and Environmental Engineering at the University of Michigan, Ann Arbor, USA. His goal as a researcher is to lay a solid foundation for human-robot collaboration by establishing safe and cohesive teaming among workers and robots. His e-mail address is civdaeho.kim@utoronto.ca. His website is https://civmin.utoronto.ca/home/about-us/directory/professors/daeho-kim/.

SANGHYUN LEE is a professor and John L. Tishman faculty scholar at the University of Michigan, Ann Arbor, USA. He earned his Ph.D. in the Department of Civil and Environmental Engineering at MIT. His research interest centers around anthropocentric construction and infrastructure management to achieve maximum performance from technologies like wearables, robotics, and automation for human. His e-mail address is shdpm@umich.edu. His website is https://dpm.engin.umich.edu/.