Semantic Relation Detection between Construction Entities to Support Safe Human-Robot Collaboration in Construction

Daeho Kim[1]; Ankit Goyal[2]; Alejandro Newell[3]; SangHyun Lee[4]; Jia Deng[5]; and Vineet R. Kamat[6]

[1]Ph.D. Student, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI 48109. E-mail: daeho@umich.edu
[2]Ph.D. Student, Dept. of Computer Science, Princeton Univ., Princeton, NJ 08544. E-mail: agoyal.cs.princeton@gmail.com
[3]Ph.D. Student, Dept. of Computer Science, Princeton Univ., Princeton, NJ 08544. E-mail: alnewell@umich.edu
[4]Associate Professor, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI 48109. E-mail: shdpm@umich.edu
[5]Assistant Professor, Dept. of Computer Science, Princeton Univ., Princeton, NJ 08544. E-mail: jiadeng@cs.princeton.edu
[6]Professor, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI 48109. E-mail: vkamat@umich.edu

ABSTRACT

Construction robots have drawn increased attention as a potential means of improving construction safety and productivity. However, it is still challenging to ensure safe human-robot collaboration on dynamic and unstructured construction workspaces. On construction sites, multiple entities dynamically collaborate with each other and the situational context between them evolves continually. Construction robots must therefore be equipped to visually understand the scene's contexts (i.e., semantic relations to surrounding entities), thereby safely collaborating with humans, as a human vision system does. Toward this end, this study builds a unique deep neural network architecture and develops a construction-specialized model by experimenting multiple fine-tuning scenarios. Also, this study evaluates its performance on real construction operations data in order to examine its potential toward real-world applications. The results showed the promising performance of the tuned model: the recall@5 on training and validation dataset reached 92% and 67%, respectively. The proposed method, which supports construction co-robots with the holistic scene understanding, is expected to contribute to promoting safer human-robot collaboration in construction.

INTRODUCTION

Autonomous robots have drawn increased attention in construction industry as an effective means of relieving human workers from the unsafe, repetitive, and unpleasant tasks of construction operations. Recently, a variety of construction robots are under development and in the early stage of deployment, including a 3D-printing robot (Zhang et al. 2018), autonomous vehicles (Sutter et al. 2018), and a humanoid robot (Kurien et al. 2018). It is expected that the successful deployment of construction robots will significantly contribute to improving both construction safety and productivity (Feng et al. 2015; Lundeen et al. 2017).

Despite such promises, persistent challenges in the co-robots' deployment have thwarted the safe collaboration between human and robot co-workers. Construction takes place in a highly dynamic and unstructured environment where multiple machines/robots and human workers

collaborate with each other in complex ways. The situational context between them evolves continually as the project proceeds. Construction robots must thus be able to understand the evolving scene contexts (i.e., semantic relations to surrounding entities), thereby safely collaborating with humans, as a human vision system does.

With the use of deep neural network (DNN) and computer vision, several construction studies have accomplished construction scene understandings, such as construction resources detection (Zhu et al. 2017; Yuan et al. 2017; Fang et al. 2018), workers' action recognition (Ding et al. 2018), and proximity monitoring (Kim et al. 2019). However, little research attempts to address the situational context understanding, such as semantic relation detection (e.g., an excavator *is guided by* a worker or an excavator *is not working with* a worker). In computer vision society, the semantic relation detection also remains one of the most challenging tasks (Newell and Deng 2017).

To address these challenges, this study builds a unique DNN architecture for semantic relation detection and develops a construction-specialized model that is fine-tuned to the construction-specific settings. Further, this study evaluates the developed model on real construction operations data so as to demonstrate its' potential in real-world applications.

TECHNICAL CHALLENGES IN SEMANTIC RELATION DETECTION

In recent years, computer vision society has made large strides with the advancement of DNN. "*Starting from breakthrough achievement in image classification from 2012, there is no computer vision applications that has not been affected by this paradigm shift*" (Girshick 2017). The scope of scene understanding is rapidly expanding as a variety of DNN architectures and learning algorithms are being developed.

Accordingly, many construction studies have leveraged DNNs [e.g., convolutional neural network (CNN) and recurrent neural network (RNN)] and computer vision, and accomplished several construction scene understandings, which include construction resources detection (Zhu et al. 2017; Yuan et al. 2017; Fang et al. 2018), workers' action recognition (Ding et al. 2018), and proximity monitoring (Kim et al. 2019). However, the semantic relation detection has not been tackled yet in the construction domain.

The semantic relation detection has recently garnered attention in computer vision society (Newell and Deng 2017). With the challenging nature of the task, it remains an open-ended study, leading to diverse approaches: fusing imagery and text data (Lu et al. 2016); using message-passing RNNs (Xu et al. 2017); predicting over triplets of object proposals (Li et al. 2017); using reinforcement learning to predict over object proposals (Liang et al. 2017). Most previous approaches depend on bounding boxes proposed from region proposal network (RPN). The use of RPN helps to break the task down into more manageable steps (i.e., two-stage inference: region proposal and semantic relation detection). However, *"this breakdown often restricts the visual features used in later steps and limits reasoning over the full contents of the image"* (Newell and Deng 2017). The separation can make an architecture not only lose the advantage of end-to-end training, but also be easily affected by errors of RPN.

In addition, developing a construction-specialized model from a DNN architecture has another challenge: how to successfully train and fine-tune the empty architecture so that it can perform well in construction-specific settings. Higher levels of scene understanding naturally demand deeper inference processing and correspondingly deeper network architecture, which in turn requires an extensive training dataset, otherwise leads to overfitting. Transfer learning offers a viable option to address this issue using a small training dataset. Pre-training with an extensive

benchmark dataset followed by fine-tuning with a relatively small construction-specific dataset will help to customize a model to construction-specific settings without the issue of overfitting. However, the fine-tuning still requires a certain amount of construction-specific dataset as well as experiments on various fine-tuning scenarios.

RESEARCH OBJECTIVE AND FRAMEWORK

To address these challenges, this study builds a unique DNN architecture (Px2Graph, Newell and Deng 2017) in which scene understandings not only for individual entities (i.e., location) but also for their semantic relations can be interactively drawn. Further, a construction-specialized model is developed by experimenting multiple fine-tuning scenarios, and validated on real construction operations data so as to demonstrate its potential in real-world applications. This study follows the below framework to achieve these aims (Figure 1).

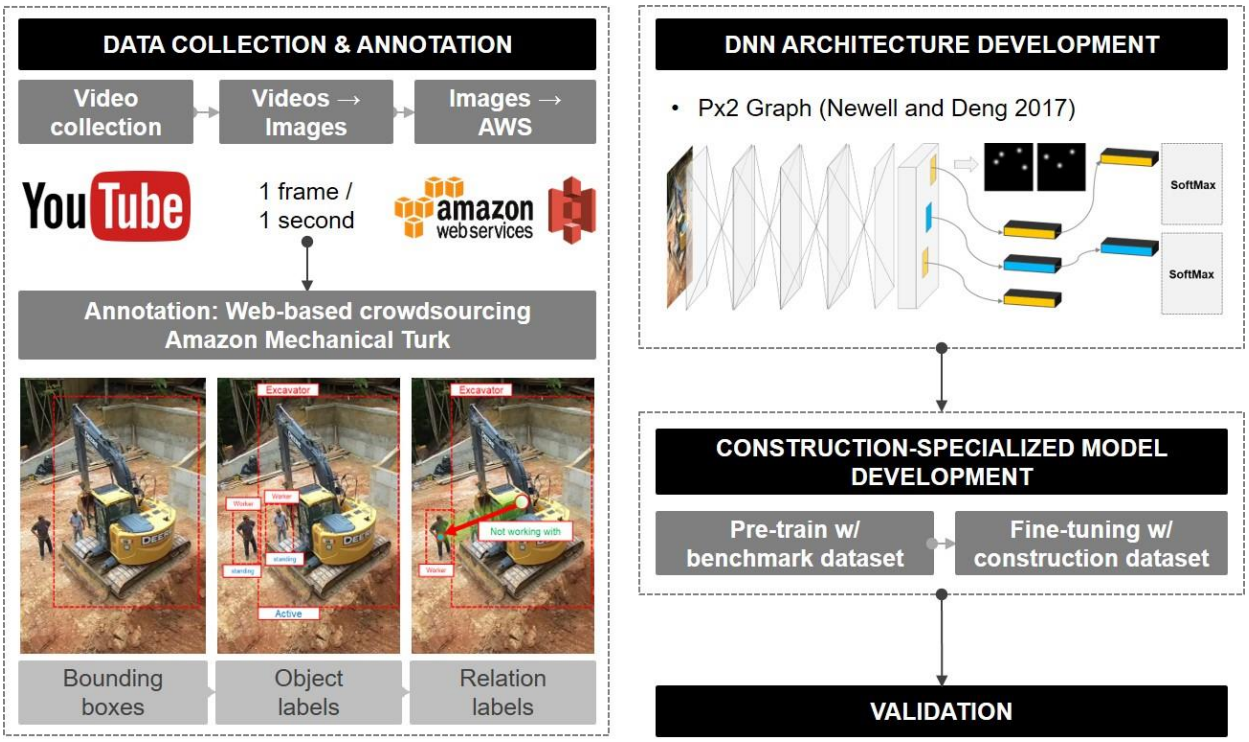

Figure 1. Research framework.

- DNN architecture development: This study builds a unique DNN architecture that can synchronously detect multiple objects and their semantic relations, leveraging hourglass networks and 1x1 convolution (Px2Graph, Newell and Deng 2017).
- Data collection and annotation: Extensive construction operations data (i.e., videos) has been collected via YouTube and annotated through a web-based crowdsourcing [i.e., Amazon Mechanical Turk (AMT)] with a complete inspection.
- Construction-specialized model development: Construction-specialized model is then developed by pre-training the proposed architecture (i.e., Px2Graph) with benchmark dataset [i.e., Visual Genome (Krishna et al. 2016)] and fine-tuning with the collected construction dataset.
- Validation on real construction data: Evaluation on real construction operations data is

conducted. Lastly, discussion on the results and implications is followed.

DNN ARCHITECTURE DEVELOPMENT

To develop an end-to-end model that can concurrently detect both object and relation, this study builds a unique network architecture that can address the following questions: (i) how to extract global features that can likely be effective for semantic relation detection and (ii) how to localize both object and relation in a single network without region proposal. The model architecture is detailed, as (Figure 2, for more information, please refer to the previous work, Newell and Deng 2017):

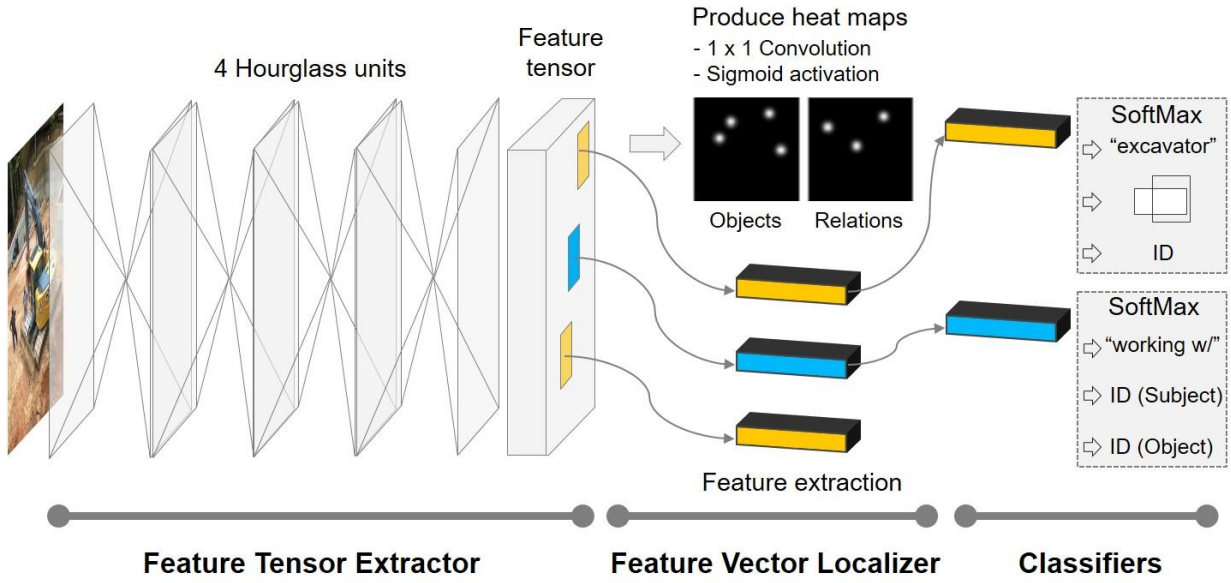

Figure 2. Network architecture: Px2Graph, Newell and Deng 2017.

- Feature tensor extractor: The four hourglass units stacked in a row takes an image as input and produces a feature tensor of fixed size. The unique design of hourglass allows the combination of global and local information, which can likely be effective in inferencing the semantic relations on a frame (Newell and Deng 2017).
- Feature vector localizer: The output tensor is then converted to heat-maps by 1x1 convolution and sigmoid activation. Each heat value represents the likelihood that an entity (i.e., object or relation) exists at the given location. The feature vectors of interest are extracted based on these likelihood values.
- Classifier: In succession, the corresponding feature vectors are fed into the fully connected layer and Soft-Max classifier, in which final classification of (i) subject class (e.g., an excavator); relation (e.g., is guided by); and (iii) object class (e.g., a worker) are performed.

DATA COLLECTION AND ANNOTATION

As an axiom of deep learning in computer vision, the quantity and quality of training dataset have a significant impact on a model's final performance. Hence, this study attempts to collect an extensive data (i.e., videos) for real construction operations and conducted frame-wise annotations with a complete inspection. First, a variety of videos for real construction sites were

collected from YouTube, which includes various scenes of human-machine (replacement of co-robots, e.g., excavator, wheel loader, or truck) interactions. Further, the authors developed an annotation template that links the collected data to web-based crowdsourcing (i.e., AMT) to reduce the avoidable efforts for massive annotations. The template leads workers to annotate each object's bounding box and relations with others (Figure 1, data collection and annotation). Lastly, manual inspection ensured the validity of the annotations. The annotation examples are illustrated in Figure 3.

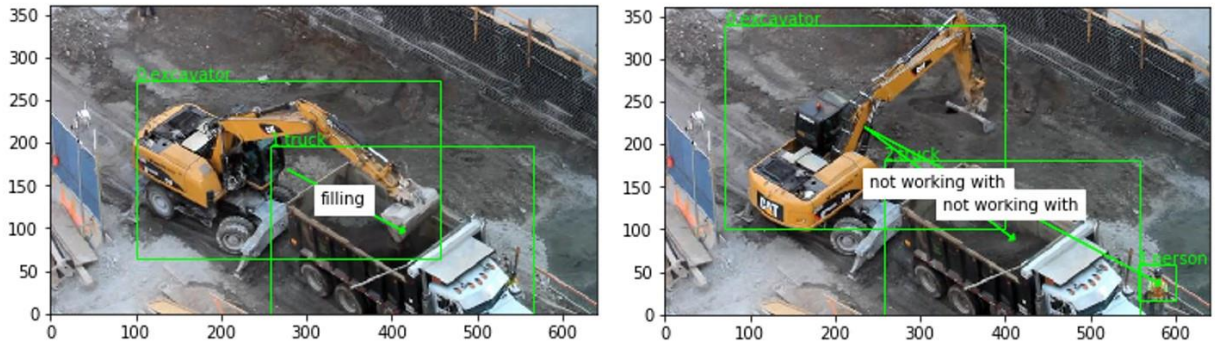

Figure 3. Examples of data annotation.

76 videos from different projects were collected. These videos capture (i) 7 types of objects (i.e., worker, excavator, truck, wheel loader, roller, grader, and van/car) and (ii) 4 types of relations (i.e., not working with, guided by, adjusted by, and filling) (Table 1). To avoid duplications in the dataset, one frame per a second was sampled in each video. As the result, the total of 2,502 frames were annotated as well as manually inspected (Table 1). This dataset includes (i) 5,468 objects and (ii) 3,110 relations in total (Table 1).

Table 1. Summary of data.

Category	Detail
The # of videos collected	76
The # of images annotated	2,502
Object categories	excavator, person, truck, wheel loader, roller, grader, van/car
Relation categories	not working with, guided by, adjusted by, filling
The # of objects annotated	5,468
The # of relations annotated	3,110

CONSTRUCTION-SPECIALIZED MODEL DEVELOPMENT

This study elected transfer learning in developing a construction-specialized model, thereby complementing insufficiency of training dataset. First, the whole network (i.e., Px2Graph) was pre-trained with Visual Genome dataset (Krishna et al. 2016) that is the most extensive dataset widely used in relation detection studies (Newell and Deng 2017; Xu et al. 2017; Lu et al. 2016). The Visual Genome contains 108,077 frames including 3.8 million objects and 2.3 million relations (Krishana et al. 2017). In succession, the fine-tuning followed with collected construction data. 2,000 (80% of total) and 502 (20%) images were used for fine-tuning and validation, respectively.

To discover a better way to transfer the pre-trained network to construction-specific settings, the fine-tuning particularly considered four different scenarios such that each scenario has

distinctive set of layers (i.e., hourglass unit) to be fine-tuned. Table 2 illustrates the four different tuning scenarios. For example, the scenario #1 fine-tunes only the last hourglass unit (i.e., 4th hourglass in the feature tensor extractor) by having zero learning rate at the other three units, whereas the scenario #4 fine-tunes all hourglass units in the feature tensor extractor.

Table 2. Fine-tuning scenarios.

Scenarios	Hourglass unit to be fine-tuned			
	Hourglass #1	Hourglass #2	Hourglass #3	Hourglass #4
S #1	X	X	X	O
S #2	X	X	O	O
S #3	X	O	O	O
S #4	O	O	O	O

Table 3. Validation results: Recall@5 of each scenario.

Scenarios	Recall @5 (%)	
	Training dataset	Validation dataset
S #1	87.78	63.90
S #2	92.20	61.68
S #3	93.62	65.12
S #4	91.93	67.41

VALIDATION ON REAL CONSTRUCTION DATA

To examine feasibility of the developed model in real-world applications, evaluation on real construction data is conducted. As an evaluation metric, this study applied recall@*x*, which is the most common metric used in relation detection studies (Newell and Deng 2017; Xu et al. 2017; Lu et al. 2016). Note that the recall@*x* reports the fraction of ground truth tuples to appear in a set of top *x* predictions. Considering diversity of the construction dataset, this work applied recall@5. The results for the four scenarios are summarized in Table 3, and graph for the recall@5 values during training is illustrated in Figure 4 with prediction examples.

It turned out that Scenario #4 (i.e., fine-tuning the entire feature tensor extractor, hourglass #1~4) outperformed all the other scenarios (Table 3). The construction dataset is highly distinctive to the Visual Genome dataset to cover universal objects and relations. Accordingly, fine-tuning the entire feature tensor extractor offered a better option over focusing on last several layers, as shown in this evaluation.

During the fine-tuning, the relation recall@5 values for training dataset steadily increased (Figure 4). Consequently, the values converged to around 90% for all scenarios (Table 3). The stably increasing pattern of relation recall@5 for training dataset shows that the proposed architecture is capable of being specialized to construction-specific settings. On the other hand, the relation recall@5 for validation dataset plateaued at around 61~67% (Table 3 and Figure 4). Although the relation recall@5 on validation dataset for all scenarios showed steadily increasing pattern, they started to converge at the early stage of fine-tuning. It is analyzed that the all scenarios suffered from insufficiency of the fine-tuning dataset, and therefore resulted in the significant overfitting.

Although the developed model showed promising performance on training dataset (i.e., more than 87% recall@5), it failed at generalization, resulting in the poor performance on validation dataset (i.e., less than 68% recall@5). It may not be sufficient for real-world applications.

However, it is noteworthy that the proposed architecture demonstrated its potential of being specialized to construction-specific settings. A follow-up study will therefore more focus on improving the generalization capability, which can include (i) augmentation of fine-tuning dataset and (ii) hyper-parameter tuning (e.g., width, height, and depth of feature tensor extractor).

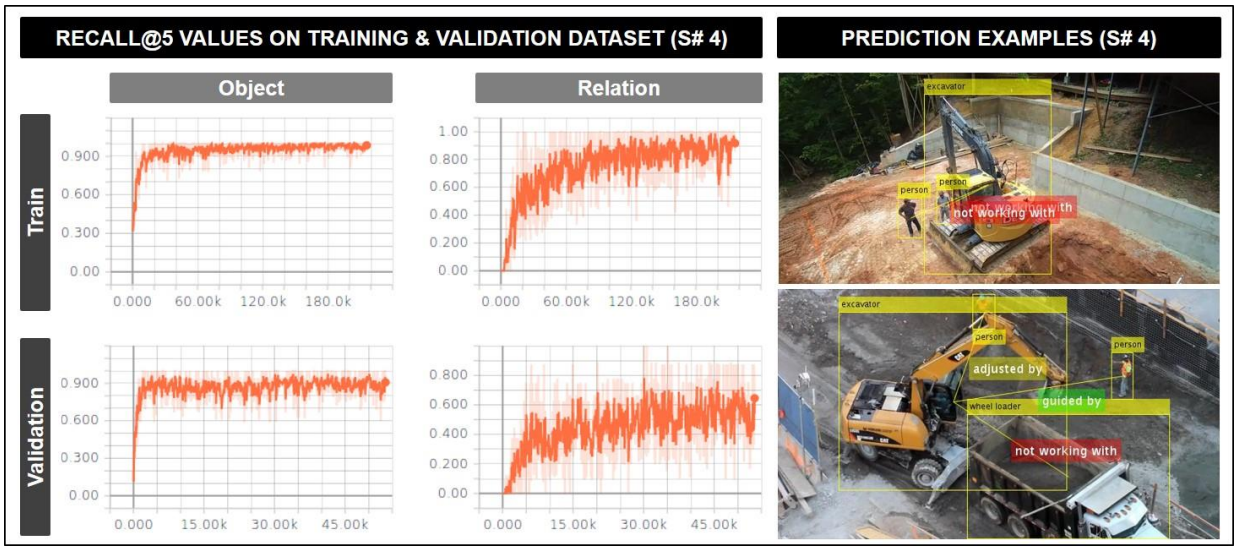

Figure 4. Results of S #4: Recall@5 during fine-tuning and prediction examples.

CONCLUSION

To support safe human-robot collaboration in construction sites, this study proposes a DNN-based computer vision method for semantic relation detection. A unique DNN architecture that can interactively detect both objects and relations is built using hourglass networks and 1x1 convolution. Further, a construction-specialized model is developed by experimenting multiple fine-tuning scenarios. As a result, the best model (i.e., scenario #4) can achieve recall@5 of 91.93% and 67.41% on training and validation dataset, respectively. The performance on validation dataset may not be sufficient for real-world applications; however, there are still plenty of opportunities to improve the performance, which include (i) augmentation of fine-tuning dataset and (ii) hyper-parameter tuning. With such critical refinement, the proposed architecture can likely result in a more robust model for construction-specific settings. The improved model will help construction robots to understand evolving scene contexts (i.e., semantic relations to surrounding entities), and it will ultimately contribute to promoting safe collaboration between human and robot co-workers in construction.

ACKNOWLEDGEMENT

The work presented in this paper 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 Environment*'). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Ding, L., Fang, W., Luo, H., Love, P.E.D., Zhong, B., and Ouyang, X. (2018). "A deep hybrid learning model to detect unsafe behavior: Integrating convolutional neural networks and long short-term memory." *Automation in Construction*, 86(2018), 118-124.

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., and Rose, T.M. (2018). "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos." *Automation in Construction*, 85(2018), 1-9.

Girshick, R. (2017). "Editorial-Deep learning for computer vision." *Computer Vision and Image Understanding*, 164(2017), 1-2.

Kim, D., Liu, M., Lee, S.H., and Kamat, V.R. (2019). "Remote proximity monitoring between mobile construction resources using camera-mounted UAVs." *Automation in Construction, 99(2019), 168*-182.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., and Fei-Fei, L. (2017). "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International Journal of Computer Vision*, 123.1(2017, 32-73.

Kurien, M., Kim, M.K., Kopsida, M., and Brilakis, I. (2018). "Real-time simulation of construction workers using combined human body and hand tracking for robotic construction worker system." *Automation in Construction*, 86(2018), 125-137.

Li, Y., Quyang, W., and Wang, X. (2017). "Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationsip detection." arXiv:1702.07191.

Liang, X., Lee, L., Xing, E.P. (2017). "Deep variation-structured reinforcement learning for visual relationship and attribute detection." arXiv:1703.03054, 2017.

Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. (2016). "Visual relationship detection with language priors." *European Conference on Computer Vision*, 852-869.

Lundeen, K.M., Kamat, V.R., Menassa, C.C., and McGee, W. (2017). "Scene understanding for adaptive manipulation in robotized construction work." *Automation in Construction*, 82(2017), 16-30.

Newell, A. and Deng, J. (2017). "Pixels to graphs by associative embedding." Advances in Neural Information Processing Systems, 2171-2180.

Sutter, B., Leleve, A., Pharm, M.T., Gouin, O., Jupille, N., Kuhn, M., Lule, P., Michaud, P., and Remy, P. (2018). "A semi-automated mobile robot for bridge inspection." *Automation in Construction*, 91(2018), 111-119.

Xu, D., Zhu, Y., Choy, C.B., and Fei-Fei, L. (2017). "Scene graph generation by iterative message passing." *IEEE Conference on Computer Vision and Pattern Recognition*.

Yuan, C., Li, S., Cai, H. (2016). "Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes." *Journal of Computing in Civil Engineering, ASCE*, 31(1), 04016038.

Zhang, X., Li, M., Lim, J.H., Weng, Y., Tay, Y.W.D., Pham, H. (2018). "Large-scale 3D printing by a team of mobile robots." *Automation in Construction*, 95(2018), 98-106.

Zhu, Z., Ren, X., and Chen, Z. (2017). "Integrated detection and tracking of workforce and equipment from construction jobsite videos." *Automation in Construction*, 81(2017), 161-171.