Toward Intelligent Agents to Detect Work Pieces and Processes in Modular Construction: An Approach to Generate Synthetic Training Data

Keundeok Park[1] and Semiha Ergan, A.M.ASCE[2]

[1]Ph.D. Student, Dept. of Civil and Urban Engineering, New York Univ., Brooklyn, NY.
Email: kp2393@nyu.edu
[2]Associate Professor, Dept. of Civil and Urban Engineering, New York Univ., Brooklyn, NY (corresponding author). ORCID: https://orcid.org/0000-0003-0496-7019.
Email: semiha@nyu.edu

ABSTRACT

Modular construction has been an alternative to traditional construction processes to reduce environmental impact and construction waste as well as to deal with space constraints in highly dense urban construction sites. Furthermore, since modules are pre-fabricated in a controlled environment, modular construction has the advantage to achieve automation and optimization as compared to traditional construction. However, due to the one-of-a-type nature of construction projects, automation in construction is still in its infancy as compared to other manufacturing industries. Meanwhile, recently, advancements in technologies such as computer vision and deep learning provide opportunities to train machine intelligence to solve problems that were not possible before. In this study, we propose an approach to automatically generate high-resolution synthetic training data for scene understanding in the modular construction context. Evaluation of the approach in testbed factory settings shows that we can systematically capture and label AEC components such as walls and doors on RGB-D images as synthetic datasets for applications of supervised learning in relation to modular construction. The proposed method can provide a mechanism to feed the necessary but missing large-scale datasets to train scene understanding models in modular construction factories as modular projects and corresponding workpieces change.

INTRODUCTION

Traditional construction is facing challenges such as the shortage of skilled workers and increasingly tight construction sites at urban settings. To alleviate such problems and bring an alternative business model to the Architectural/Engineering/Construction (AEC) industry, prefabrication and modular construction are gaining wider acceptance. The main working principle of modular construction is to preassemble modules (i.e., volumetric units or panels) and get them stacked and integrated at construction sites. Since module prefabrication is processed in a controlled indoor environment, it provides increased productivity and a better environment for quality control. In addition, modular construction has advantages in environmental aspects. Modular construction has benefits in the reduction of construction waste by 15% and also in lower impacts on surroundings by 50% (Lawson et al., 2012). Given that the impact of construction projects on surroundings during construction is inevitable, it has significant benefits, especially in dense urban areas.

Similar challenges that we see in a traditional construction project are also observed in modular construction projects. Unique nature of design and configurations result in ever-changing work

pieces, panel, and volumetric unit configurations that could hinder productivity and efficiency while locating and identifying related pieces. Recent developments in AI and robotics provide opportunities in human- robot collaborations to enhance productivity, efficiency, and throughput at production yards. Although earlier efforts on robotics and automation dates back early 1980s (Bock, 2015), main applications were for robots to perform physical tasks instead of utilizing them for their robust information processing capabilities as intelligent agents. Such intelligent agents, however, need to be trained using large scale structured data on scenes, with distinct geometries, assembly stages, and positionings of assemblies/work pieces. Importance of the quality of data used to train such intelligent agents is well-known (Ng, 2021). Hence, there are various efforts across the domains to curate high-quality data and benchmarking mechanisms for successfully developing agents. The efforts on autonomous vehicles that provide Embodied AI and learning environments for training self-driving AI are notable examples. However, the AEC industry and applications that could benefit from intelligent agent lack such platforms to provide the massive data needed in the training of such agents. This work addresses this need for platforms and mechanisms to curate and generate the datasets needed. Within the modular construction context, this approach utilizes virtual environments to enable dataset generation on scenes that are needed to train embodied AI for recognizing workpieces and steps in assembling components at factory settings. Embodied AI is an emerging field in AI for agents to learn thorough physical interactions with their environments in simulated in virtual worlds that provide task-based datasets instead of large distinct datasets (e.g., images, videos) that are used in traditional AI (Smith, 2005; Savva et al., 2019). Since it is time-consuming, labor-intensive, and costly to collect such task-based datasets to train agents/robots, embodied AI simulators ease this data curation process. This paper provides an overview of embodied AI based data curation process along with its evaluation in modular factory settings.

BACKGROUND

Approaches for training intelligent agents can be separated into two main categories as: (1) training models using well-structured 3D datasets with rich scene information; (2) developing embodied AI simulators for virtual robots.

Utilizing well-structured 3D datasets to facilitate AI training with rich scene information

It is the quality and quantity of data that determines the performance of AI models. Nowadays, the computer vision domain provides pretrained models that utilize large-scale genetic image datasets ImageNet; MS-COCO) for various application domains. The objective of 3D datasets is to provide a 3D environment similar to the physical world, in which AI can gain intelligence for scene understanding. In other words, the environment in a studied domain should provide rich information to agents for gaining domain specific intelligence.

The environment can be indoors or outdoors as per the context. For a smooth learning workflow, data should be well-structured with RGB-D (Red, Green, Blue and Depth), annotations on objects in scenes (e.g., 2D-3D bounding boxes and classification labels, semantic labels), relationships between objects in scenes (e.g., scene graph generation from objects) (Armeni et al., 2019). Representative datasets used so far include AI2-THOR (Kolve et al., 2017), which includes virtual indoor environments from real environment of 89 apartments with more than 600 objects and generated initially for training robot AI to navigate and conduct tasks such as opening a door

and bringing a laptop within household context; Matterport3D (Chang et al., 2017), which has 194,400 RGB-D indoor images captured via 3D camera from 90 real buildings scenes that were originally used for providing datasets for researchers on scene understanding; and Gibson env (Xia et al., 2018), which includes reconstructed scenes of the entire building from 572 buildings through conversions of raw point cloud data, where scenes have semantic depth, labels of objects and normals of faces and used originally for training AI to navigate 3D spaces by giving real-world perception. Among these, AI2-THOR and Gibson datasets are generated using platforms with embodied agents. An overview of some of these 3D datasets generated and provided for research communities is provided in Table 1 below.

Table 1. Examples of structured 3D Datasets

Dataset name	Scene/environment data captured	Dataset description	Source of data
SUN3D	Indoor (household)	Labeled images, point clouds	Real captured with a depth camera
ScanNet	Indoor (household)	Point clouds	Real captured with a depth camera
AI2-THOR	Indoor (household)	3D Environment	Real, Synthetic captured with a camera and built in Unity
Structured3D	Indoor (household)	Labeled meshes and RGB-D	Synthetic generated with 3D graphics software
3D-FRONT	Indoor (household)	Labeled meshes	Synthetic generated with 3D graphics software
Matterport3D	Indoor (household, office, church)	Labeled point clouds	Real captured with a depth camera
Gibson	Indoor (household, office)	Labeled point clouds	Real captured with a laser scanner and a depth camera

These datasets are publicly available for research purposes and have been utilized mainly for benchmarking for 3D reconstruction methods (Liu et al., 2018; Dai et al., 2021) and scene understanding (Armeni et al., 2019; Dai et al., 2018; Liu et al., 2018). Such datasets helped to verify 3D reconstruction of scenes using deep learning methods, semantic segmentation (Dai et al., 2018), prediction of occluded regions in scenes (Liu et al., 2018), and spatial relationships represented in 3D graphs between objects (e.g., a refrigerator is in the kitchen) (Armeni et al., 2019). The generation of these datasets also shows differences, as can be seen in Table 1, with unique limitations. They are either real world 3D scene data, where scenes have been scanned with LIDAR technology and then reconstructed in 3D; or they are synthetic data generated by 3D computer graphic tools with rendering techniques such as 3D Max, Blender, Sketchup, Unity, Unreal, and Maya. Capturing real scenes requires robust real world scene capturing technologies and diverse set of environments/scenes to reflect as many environments and tasks as possible, whereas synthetically capturing scenes requires exact replica of the actual scenes (e.g., an indoor household unit design complying with the building code and/or real dimensions captured) and their realistic appearance to minimize the scene gap. Both methods have been actively used for dataset generation by alleviating the limitations of each.

Embodied AI simulators for intelligent agents to interact with 3D environments

Embodied AI is an emerging field in AI for agents to learn through physical interactions with their environments in simulated virtual worlds that provide task-based datasets instead of large distinct datasets (e.g., images, videos) that are used in traditional AI (Smith, 2005; Savva et al., 2019). Distinct components of an embodied AI simulator include the types of agents being trained (e.g., Humanoid, Ant, Robot dog), context dependent tasks (e.g., navigating by sounds, find an object and bring back, open a door), and environments agents operate. The benefits of using embodied AI simulators to train agents are many and include (a) *a fast learning process,* where parallel learning that simultaneously executes multiple AI training sessions is possible; (b) *increased safety* that reduces the injury of surrounding and people when errors occur in learning of the real robots; (c) *feasibility*, where physical robots and diverse environments are expensive to put together; and (d) *reproducibility*, where published environment and AI can be reproduced so that research communities can continuously improve the intended technology (Savva et al., 2019).

Current leading efforts on embodied AI simulations include Nvidia's ISAAC Sim, which has been developed for the application of NVIDIA's robot SDK, Facebook Research's Habitat-sim, which aims to build a platform with configurable agents and AI algorithms connecting existing 3D datasets (i.e., Matterport3D and Gibson env) for training AI algorithm with different types of robot bodies in various environments (e.g., different types of buildings) virtually, and AI2-THOR, which provides a specific agent with 3D environment they build to teach visual AI to conduct ordinary tasks in a household context. Simulators provide various types of agents that reason with the scene and perform tasks. For example, the virtual robot model in AI2-THOR has a 1.3m height mobile body and has a 6 degree of freedom manipulator, which is the general design of an articulated robot arm, to perform tasks such as grabbing a cup and placing it in the sink. In general, these simulators provide a rich context for basic tasks such as the navigation of robot agents (e.g., navigating to a set destination, to the source of a sound, to a specified object, etc.) and interactions of agents with objects in these virtual environments.

In summary, structured 3D data and embodied AI simulators are complementary to each other. For instance, recently published structured 3D datasets are possible to be imported in embodied AI simulators (e.g., Matterport3D and Gibson env in Habitat-sim). Since setting camera pose freely in 3D coordinates is essential to apply embodied AI simulators, previous scene understanding datasets (e.g., MS-COCO), which provide only 2D perspective on labeled images, are limited in this domain.

The study presented in this paper leverages the strengths of both approaches by valuing the importance of having well-structured 3D datasets in training human collaborators as agents, and by utilizing simulators to eliminate the burden to manually capture such datasets in different factory and modular construction project settings. The approach helps to populate a well-structured and realistic 3D dataset on modular construction scenes, including related objects in modular factories (e.g., overhead gantry cranes, forklifts, building components), simulated assembly tasks, and assembly sequences that are captured systematically. Generation of such datasets will provide the much-needed datasets for training embodied AI in modular construction factory context along with the generated 3D dataset.

An Embodied-AI Based Virtual Simulator to Curate a Dataset for Training Robot Agents for Modular Construction

Scene understanding refers to the ability of machine intelligence to visually analyze a scene to interpret the current situation (e.g., what objects are appearing in the scene) and able to find

adequate action if it is asked (Hoiem et al., 2015). This work focuses on populating realistic 3D datasets for scene understanding models as an initial step towards training intelligent agents to learn modular construction factory environments and tasks to assist human workers in work piece identification and locating them with respect to upcoming assembly steps. In a nutshell, scene understanding is a process that is based on supervised learning to perform tasks such as object detection (e.g., detection of 80 unique objects such as person, chair, and sofa using the COCO dataset), semantic segmentation (e.g., labeling 33 unique objects such as cars, fences, pedestrians, and bikes mostly relevant to the street from the perspective of a driver using the Cityscape dataset), and object pose estimation (e.g., inferring 3D representation from 2D perspective images using T-LESS dataset). Earlier scene understanding models were trained using real photos that were taken and labeled manually. Since collecting a large-scale data related to a specific problem domain is tedious and time-consuming, models built off of generic large-scale datasets (e.g., ImageNET) have been used as baselines and have been extended to other application domains eliminating the effort to train models from scratch. In recent years, advancements in computer graphics provide photorealistic texturing and rendering that fill the gap between the appearance of reality and 3D graphics, which can be leveraged to generate diverse realistic synthetic datasets with automatic labeling within virtual environments. This approach builds on these advancements.

In a nutshell, the workflow of the developed approach to generate labeled image sets from modular construction factory settings is as follows (Figure 1): 1) build the virtual factory environment using LIDAR data, 3D graphics tools such as Blender & Unity; 2) automatically disaggregate digital volumetric units into individual module work pieces (e.g., chasses, window panels, wall panels); 3) place disaggregated module components into the virtual factory environment, 4) simulate the assembly process, and 5) set cameras systematically to capture RGB-D images and labels as the assembly is in progress. For high quality virtual environments, the first step of building the environment could require manual work depending on how much data is already digitally available about the manufacturing yard, and subsequent steps could be done automatically or semi-automatically through BIM authoring tool's API and parametric configurations in Unity. The outcome of this process will be a set of (a) RGB images, (b) depth images, and (c) corresponding labeled images (color masks) captured systematically during the assembly of volumetric units that will be needed for scene understanding. Details of the steps are provided next.

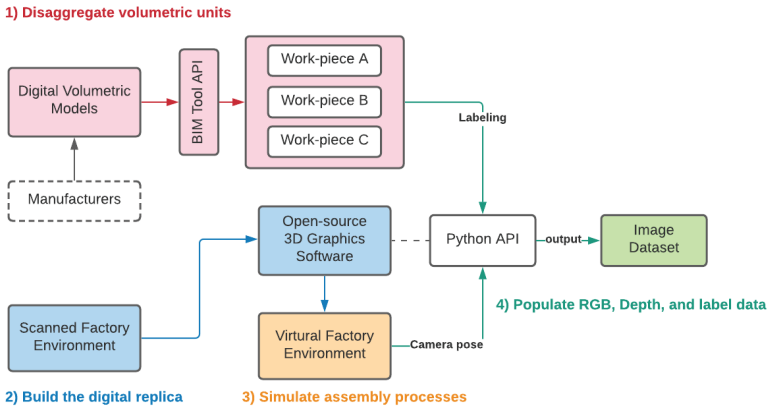

Figure 1. Overview of building virtual simulators to populate labeled image sets

Step 1: Disaggregate digital volumetric units into individual module work pieces

A work piece could be a single element such as a column, beam piece, or it could be a preassembled panel (e.g., structural wall panel) that could be part of a volumetric unit (Figure 2). Initial building information models of volumetric units have been obtained from opensource 3D warehouses (BIMOBJECT), where manufacturers upload their design and products for marketing purposes. Modules were in Revit file format, which contained unique volumetric units or preassembled wall panels (Figure 2). First, since there is no built-in function to save each component as individual IFC files in a BIM authoring tool, we first separated volumetric units and panels into individual work pieces using BIM authoring tool's Python API. We implemented a code that takes 3D models of a set of work pieces with unique IDs that were assembled as a volumetric unit and transforms them into individual 3D models. Each work piece has a specific family where subcomponents/parts (if any) belong to that geometry and grouped together. Following are the actions that the code does when executed: 1) generate a list of components in the given model using their IDs, 2) select a work piece (e.g., a wall panel type A) from the 3D model and record the ID of the selected piece; 3) remove the other components except the selected component; 4) Save the model in IFC format and mark the ID of the selected component from the list before the iteration starts again; 5) iterate the process until all individual components are selected from the list and their IFC files are saved. The stored individual work pieces are shown as Figure 2.

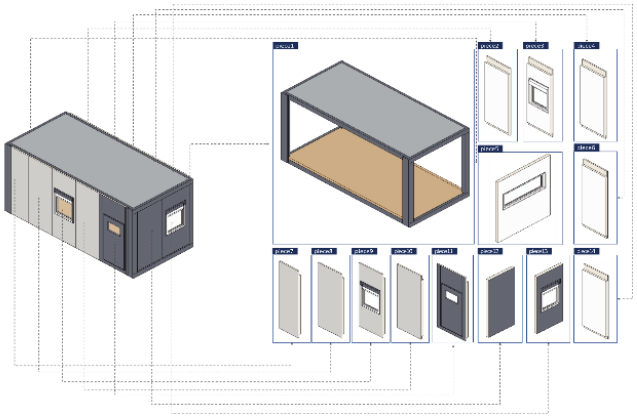

Figure 2. Disaggregation of volumetric units and panels into work pieces

Step 2: Build the digital replica of a factory environment

A digital replica of a factory layout is needed for replicating the environment the agents will operate in. This environment also requires digital models of typical objects (e.g., forklift, shelves, overhead crane, etc.) in that factory environment that the agents will interact with. Geometric accuracy and realistic representation of the virtual environment is an essential requirement in order to have high quality datasets for supervised learning. 3D environments can be generated using various ways including the 3D reconstruction techniques (i.e., Photogrammetry) or reality capture technologies (e.g., LIDAR). We have captured a factory environment by LiDAR and used the point clouds to reconstruct the environment in the virtual environment (VE). Figure 3 shows an image of the factory the team has access to, and the digital replica of it in VE.

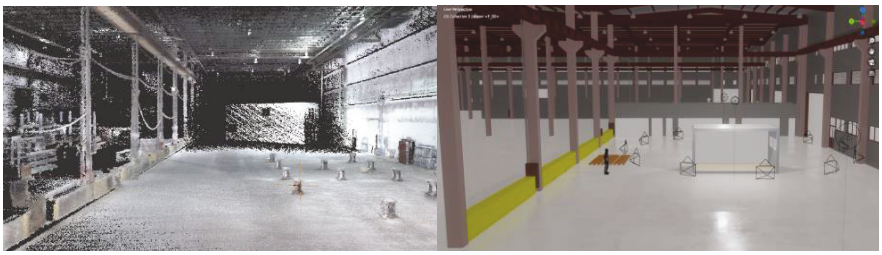

Figure 3. Scanned modular construction factory (left); Virtual factory environment in open-source 3D computer graphics software (right)

For realistic factory modeling, not only the factory layout and modules, but also other modular construction factory related objects inside the factory are required (e.g., forklift, coils, overhead cranes, shelves). In this study, we used 3D assets from an online 3D warehouse and generic objects from ISAAC Sim (e.g., forklift, trolley, shelves). 3D work pieces and volumetric units have been placed on where typically the components are placed and assembled (Figure 3 right). We plan to expand the data variation by adding module products provided by various partnering companies.

Step 3: Simulate the volumetric unit assembly processes

Inference of module fabrication progress is an essential feature of AI to help workers in the manufacturing factory towards increasing the throughput and reduction of rework. However, since the module fabrication processes differ depending on the manufacturer, the sequence of module fabrication must be defined. In this study, we used a sample module product sequence as shown in Figure 4 for testing the developed approach.

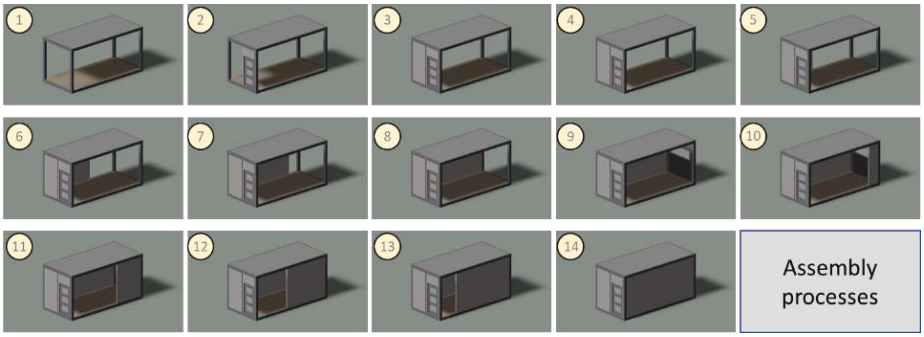

Figure 4. A sample sequence of assembly steps is implemented for testing the approach

We used predefined family components in the BIM authoring tool's API. For example, the structural frame of module is already fabricated as single object in the given model. Followings define the sequence of sample module fabrication (Figure 4): (1) preassembled structural frame is placed; (2) main entrance component is assembled; (3) wall panels are assembled to the main frame as eleven individual members in the clockwise direction starting after the main entrance installation; (4) window component is placed in the wall opening on the wall panel (wall panel on image frame # 9 in Figure 4). Dataset generation requires that this sequence is set so that the assembly progress can be captured and populated in the image set.

Step 4: Populate the RGB-D and labeled image dataset

In this study, we utilized the developed platform for the generation of RGB images, depth maps, and semantic labeled images. Various types of data (e.g., RGB, Depthmap, Normals, bounding boxes, semantic labels, point clouds) can be generated using 3D computer graphics software that allows scripting such as Blender or Unity. Within the virtual 3D environment, datasets for scene understanding can be generated using 3D authoring tools. Specifically, open-source plugins in such tools to generate image datasets have no significant difference between both platforms. We built the 3D environment based on scanned factory environment and imported individual components from disaggregated volumetric units into scenes and then implemented an algorithm to generate corresponding RGB-D and labels from virtual camera poses. Within the virtual environment, the area for module assembly was defined first. To capture the module objects, virtual cameras were placed in upright 12 viewpoints and dodecahedron 20 viewpoints format. Thereafter, RGB-D images were captured through virtual camera in the platform and labeled semantics are represented as classes of objects, which were retrieved from corresponding model elements in BIM (e.g., an object named as "Wall panel Type A" has the class name as "wall").

Previous efforts in the computer science domain utilized multi viewpoint (12-viewpoints) camera systems to capture image sets and the results show from 0.85 to 0.95 accuracy in object classification (Koo et al., 2021; Li et al., 2020). It is possible to use 12-viewpoints and additional viewpoints camera systems (Dodecahedron) to capture images systematically. Figure 5 provides the upright 12-viewpoint camera capturing system (left) and dodecahedron 20-viewpoint camera capturing system. The dodecahedron camera system has more diverse views from different angles.

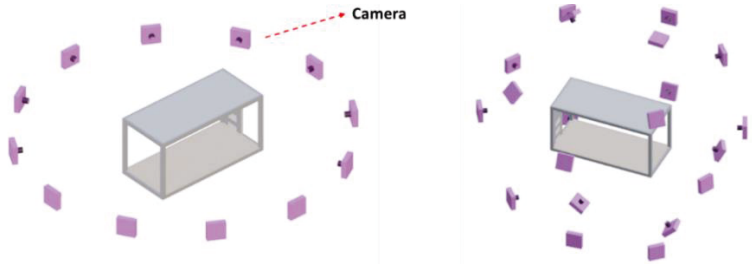

Figure 5. 12-viewpoint camera system (left); Dodecahedron-viewpoint camera system (right)

IMPLEMENTATION OF THE APPROACH AND GENERATED DATASET

The approach has been implemented using a 12-viewpoint virtual camera system. The generated dataset for each scene includes raw images (RGB), semantic labels of each component seen in an image (e.g., frame, ceiling, floor, door, window, wall), and depth map images. RGB images are normal images captured from the 12-viewpoint virtual camera system, semantic labels are the color-masked images of components, and depth images represent the distance information between the virtual camera system and work pieces in the scenes (see Figure 6). We evaluated the data generation approach in modular construction context and used the approach to generate such a dataset for training a vision-based model to identify and label the sequence in a modular volumetric assembly process (Park & Ergan, 2021). We utilized real modular construction assemblies, their assembly sequences, and a manufacturing yard to generate the virtual

environment. Using this virtual environment and the virtual camera system, we generated 7,000 sets of inputs, which include 84,000 images in total, with labels indicating the assembly sequence seen in the corresponding scenes. These 84,000 images with labels were used to train and test models to classify sequences of module assemblies. The model on the test dataset showed 0.97 overall accuracy to identify and label the unseen module assembly sequences. The evaluation of the approach and the generated dataset using this approach has been detailed in Park & Ergan (2021).

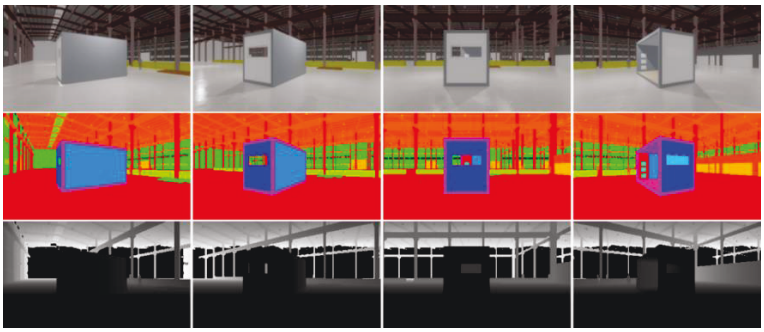

Figure 6. A glance at the dataset: RGB Images (top); Semantic labels (middle); and Depth Images (bottom) captured using the 12 multi viewpoint camera system

CONCLUSION AND FUTURE WORK

This paper provides an overview of an approach to streamline synthetic data generation for scene understanding within modular construction context. It enables to generate labeled image datasets whenever the products are changed in the factory. This approach enables creating the virtual model of any modular construction factory and generating well-structured dataset for scene understanding in the modular construction factory context. As well-structured 3D environment and dataset are essential to embodied AI for real world application, this study is the first step towards the development of AI helpers for modular construction factory workers.

The integration of the presented study and embodied AI research specialized in modular construction factory can provide an opportunity to create collaborative robots working with human workers, which has knowledge of the context of modular construction factories such as equipment used in the factory and assembly processes. As this study is focusing on the 3D environment and data generation for training AI, in order to achieve flexible and accurate AI, further steps are needed to streamline a process to change modular construction factory environments and assembly processes in virtual settings and integrate embodied AI by adopting diverse types of agents and domain specific tasks.

REFERENCES

Lawson, R. M., Ogden, R. G., and Bergin, R. (2012). Application of modular construction in high-rise buildings. *Journal of architectural engineering*, 18(2), 148-154.

Bock, T. (2015). The future of construction automation: Technological disruption and the upcoming ubiquity of robotics. *Automation in Construction*, 59, 113-121.

Ng, A. (2021). "Issue 84". Retrieved from https://www.deeplearning.ai/the-batch/issue-84/, access date: March 24, 2021.

Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review*, 25(3-4), 278-298.

Armeni, I., He, Z. Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J., and Savarese, S. (2019). 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5664-5673).

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gorden, D., Zhu, Y., Gupta, A., and Farhadi, A. (2017). Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158.

Xia, F., Zamir, A. R., He, Z., Sax, A., Malik, J., and Savarese, S. (2018). Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE on CVPR* (pp. 9068-9079).

Liu, C., Wu, J., and Furukawa, Y. (2018). Floornet: A unified framework for floorplan reconstruction from 3d scans. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 201-217).

Dai, A., Siddiqui, Y., Thies, J., Valentin, J., and Nießner, M. (2020). Spsg: Self-supervised photometric scene generation from rgb-d scans. arXiv preprint arXiv:2006.14660.

Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., and Nießner, M. (2018). Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4578-4587).

Liu, S., Hu, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X. (2018, December). See and think: Disentangling semantic scene completion. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 261-272).

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., and Batra, D. (2019). Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9339-9347).

Koo, B., Jung, R., and Yu, Y. (2021). Automatic classification of wall and door BIM element subtypes using 3D geometric deep neural networks. *Advanced Engineering Informatics*, 47, 101200.

Li, Z., Wang, H., and Li, J. (2020). Auto-MVCNN: Neural Architecture Search for Multi-view 3D Shape Recognition. arXiv preprint arXiv:2012.05493.

Hoiem, D., Hays, J., Xiao, J., and Khosla, A. (2015). Guest editorial: Scene understanding. *International Journal of Computer Vision*, 112(2), 131-132.

Park, K., and Ergan, S. (2021). Towards Intelligent Agents to Assist in Modular Construction: Evaluation of Datasets Generated in Virtual Environments for AI training. *38th International Symposium on Automation and Robotics in Construction* (submitted).