Multi-task Deep Learning-based Human Intention Prediction for Human-Robot Collaborative Assembly

Jiannan Cai, Ph.D., A.M.ASCE;[1] Xiaoyun Liang, S.M.ASCE;[2]
Bastian Wibranek, Ph.D[3], and Yuanxiong Guo, Ph.D.[4]

[1]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio (corresponding author). Email: jiannan.cai@utsa.edu
[2]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio. Email: xiaoyun.liang@utsa.edu
[3]School of Architecture & Planning, The University of Texas at San Antonio. Email: bastian.wibranek@utsa.edu
[4]Department of Information Systems and Cyber Security, The University of Texas at San Antonio. Email: yuanxiong.guo@utsa.edu

ABSTRACT
Construction robots have great potential to serve as assistants to relieve construction workers from repetitive and physically demanding tasks. It is essential for robots to understand and predict human intention in order to adapt their motion to ensure smooth human-robot collaboration. This study proposes a long short-term memory model-based multi-task learning framework to simultaneously predict multi-level human intention in assembly tasks, including high-level actions and objects, and low-level body movements, from observed body movements and associated assembly components extracted from videos. The proposed models were trained and tested using 54 videos collected with nine participants performing six assembly tasks, achieving an accuracy of 82% and 98% in action and object prediction, respectively, and an average displacement error of 8.71 pixels in pose prediction. The incorporation of work context significantly improves the accuracy of object prediction by 11.36%, with the performance of other two tasks increasing slightly.

INTRODUCTION
Robotics has attracted broad attention as an emerging technology in construction, with applications evolving over the past years from single-task construction robots, such as brick-laying robots (Madsen 2019), rebar-tying robot (Cardno 2018), to general-purpose robotic platforms, especially collaborative robots, for more flexible human-robot collaboration (HRC) (Kim et al. 2021). Collaborative robots have great potential to serve as assistants to relieve human workers from repetitive and physically demanding tasks, such as lifting and transporting heavy objects, tools handover, assembly, etc. To this end, it is essential for robots to understand and predict human intention, such that they can reason about their task and adapt their motion to ensure smooth and effective HRC.

Many studies have developed methods to predict high-level human intention in terms of their potential actions and/or desired objects in HRC, mostly in the manufacturing sector. For instance, Liu and Wang (2017) developed a Hidden Markov model to predict pre-defined groups of human actions. Liu et al. (2019) created a convolutional neural network (CNN)-long short-term memory (LSTM) framework to predict intended action and desired tool in manufacturing assembly tasks. Wang et al. (2022) proposed a method to predict human intention, represented as a group of commands (e.g., stop, continue, slow down, etc.) to robots, via multimodal signal inputs, including

natural language and wearable sensors. Some studies focused on understanding low-level movements, but many were limited to hand movement prediction for HRC. For instance, Luo and Mai (2019) developed a probabilistic dynamic movement primitive model to predict human hand motion in manufacturing task. In construction sector, Zhou et al. (2022) proposed a framework to cluster human behavior in terms of gaze-hand relationship and predict hand movement based on gaze trajectories in pipe skid maintenance task. From the review of existing studies, most research only focuses on a single type of human intention, without a holistic understanding of multi-level intention, which is critical to developing an intelligent robot that can adapt to various human behavior for smooth collaboration.

To overcome this limitation, this study proposes an integrated framework to simultaneously predict multi-level human intention in an assembly task, including high-level actions and objects, and low-level body movements, from observed body movements and associated assembly components extracted from videos. The contribution lies in two aspects: 1) a LSTM-based multi-task learning (MTL) model is developed to predict multi-level human intention, including high-level actions and objects and low-level movements, leveraging the commonality in human behavior. The encoder-decoder architecture of the model enables the prediction over multiple timesteps. Furthermore, the model is augmented with task context information, i.e., current components being assembled, and shows improved performance compared to conventional methods that only consider human movements. 2) The proposed integrated framework allows the robot to answer three critical questions simultaneously in a HRC assembly task, i.e., "what is needed?" – object prediction; "when is needed?" – action prediction; "where to pass the object?" – body movement prediction. Knowing this enhances the intelligence of assistant robots and enables them to adaptively plan their motion based on human intention. This framework can be applied to other sectors, such as manufacturing, healthcare, etc.

METHODOLOGY

The methodology consists of three steps, as shown in Figure 1. First, collaborative assembly tasks were designed, which involve different repetitive actions (e.g., pick, carry, assemble, etc.) and objects (e.g., main parts, connector). A series of experiments were conducted and recorded, and the collected videos were used for training and testing the proposed framework. Second, human body movements, represented by time-series skeleton poses, were extracted from videos using a deep learning model. Third, a LSTM-MTL model was created to predict high-level actions and objects, and low-level body movements using time-series skeleton poses from the second step. Furthermore, the objects involved in the current actions were incorporated as contextual information to improve prediction performance.

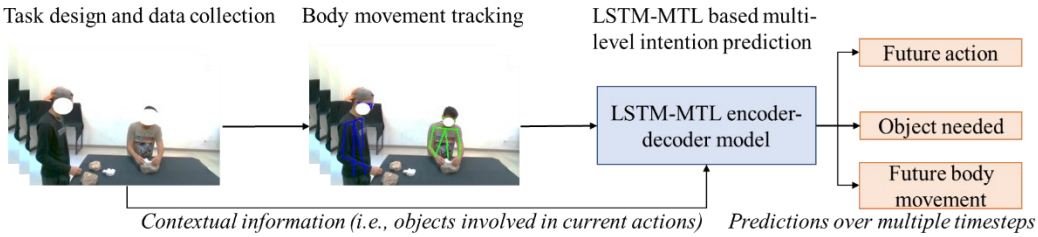


Figure 1. Overall framework

Task Design

For this study, six column-like structures were designed using unprocessed stones and 3D-printed connectors for dry-joined assembly without any adhesives. The stones were 3D scanned using

photogrammetry, and for each column, the stones were aligned along the vertical thrust line with their center of mass. 3D-printed connectors were designed and fabricated according to the arrangement of these stones (see Figure 2). These structures required high accuracy due to the perfect fit connections of the dry-stacking of the unprocessed stones and the algorithmically generated 3D printed connectors (Wibranek and Tessmann 2019). The tasks were designed with different levels of complexity, including three tasks with two stones and one connector and three tasks with three stones and two connectors (see Figure 3), requiring participants to adjust relative positions and orientations between stones and connectors to precisely align with the design. These tasks simulate complex assembly tasks in construction that require human dexterity and experience, but robots could serve as assistants to hand over components and tools based on human needs.

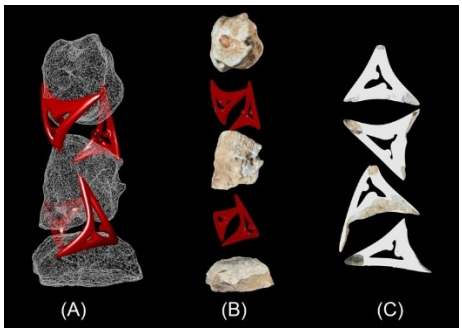


Figure 2. Digital representations, highlighting the accuracy of the parts (A), an exploded view with the three stones and 3D printed parts (B), and top views of the interfaces between stones and 3D printed parts (C)

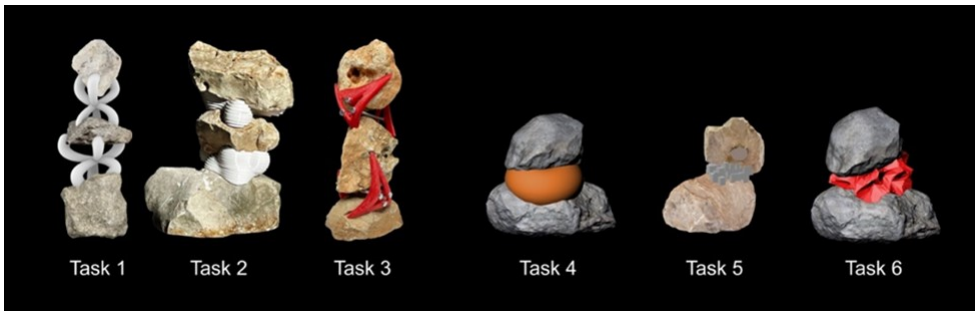


Figure 3. Assemble tasks (Tasks 1-3 are more complex than Tasks 4-6)

Nine groups of participants were recruited to perform all tasks. Each group consists of two people, with one focusing on assembly task and the other serving as an assistant. The principle is that human-robot collaboration could learn from conventions in human-human interactions (Krämer et al. 2012). Therefore, the human behavior collected during human-human collaboration in the same settings could be valuable to train deep learning models to predict human intention, which could be then embedded in robot platform to enable intelligent HRC.

The entire experiments are videorecorded, and the collected videos serve as input data for the integrated framework to predict human intention. Specifically, six types of actions that are generalized from all tasks were defined, i.e., pick, carry, assemble, adjust, inspect, and release (see Figure 4). Depending on specific task and participant, the length and order of each action may vary in each experiment. Some actions may be even absent from certain experiments, for instance, some participants may not “inspect” and/or “adjust” when assembling simple structures. Such variability

ensures the diversity and uncertainty of training data and highlights the need of deep learning-based framework to improve generalization capability.

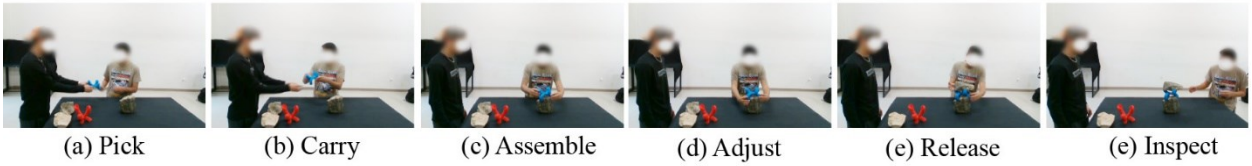


Figure 4. Samples actions involved in assembly tasks

Tracking of Human Body Movement

Low-level human movement is represented as time-series locations of human key points, such as shoulders, elbows, hands, ankles, etc. This study adopts a deep learning-based pose tracking algorithm, Pose Flow, developed by (Xiu et al. 2019), to track key point movements considering its flexibility in multi-person scenarios and the computational efficiency. Particularly, the original pose tracking algorithm tracks the movement of 17 key points for full body. In the task setting of this study, lower body is usually occluded by the work bench and the collaborative assembly mainly involve upper body movement. Therefore, only 13 key points of upper body (including hips) are considered.

LSTM-MTL Model for Intention Prediction

This study proposes an LSTM encoder-decoder model based MTL framework to predict multi-level human intention (see Figure 5). High-level action and object prediction are formulated as multi-class classification problem, while low-level movement prediction is modeled as regression problem. Specifically, the 2D pixel coordinates of 13 human key points obtained from pose tracking are normalized and constructed into a 26-dimensional feature vector, which are then concatenated with the object information to incorporate the work context, i.e., the current work status, represented by the object that is currently handled or targeted by the person. The object class contains three groups, i.e., stone, connector, and stone integrated with connector (in Task 3 only). The resulting 27-dimensional features over multiple time steps are fed into LSTM encoder that captures the dynamics of human movements and work context.

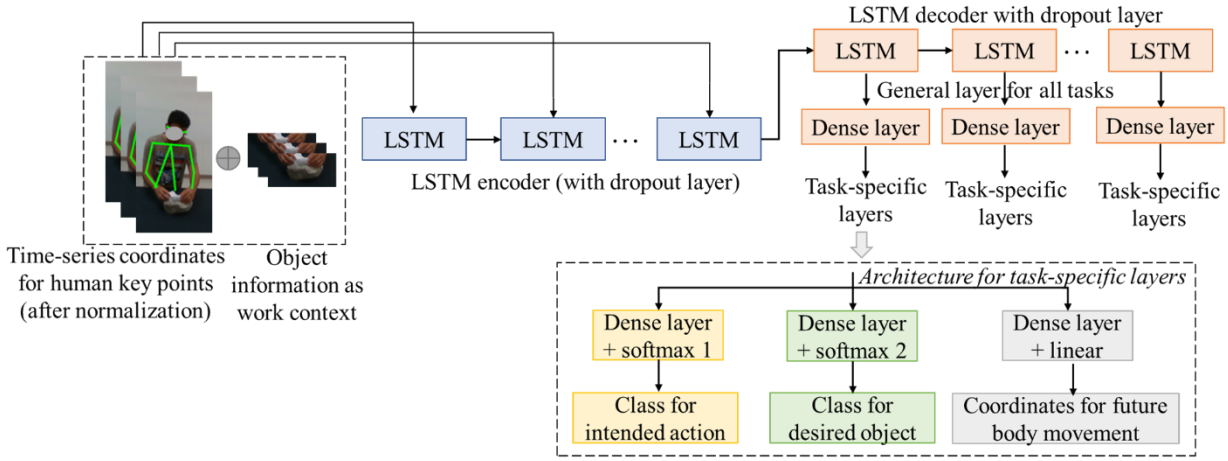


Figure 5. Proposed LSTM-MTL model for multi-level human intention prediction

MTL mechanism is adopted to simultaneously learn models to predict multi-level human intentions, as it can effectively increase computation efficiency and mitigate the challenges when specific classes have limited training data (Cai et al. 2021, 2022). Specifically, the encoded vector is fed into LSTM decoder for multi-step prediction, which is followed by a general dense layer to capture the common representation across all tasks. Then, task-specific dense layers are used to model the uniqueness of individual tasks, resulting in separate outputs for each task. Besides, dropout layers are applied for both LSTM encoder and decoder to mitigate overfitting issue. The loss function of this LSTM-MTL model is the weighted combination of categorical cross-entropy losses for action and object prediction, and mean squared error (MSE) loss for movement prediction. The weights could be determined based on the relatively importance of each task.

IMPLEMENTATION

The proposed framework was evaluated using videos collected from the experiments – a total of 54 videos were collected with 9 participants assembling 6 different structures, as detailed in "Task Design" section. All videos were taken at 15fps. The Pose Flow algorithm (Xiu et al. 2019) was first applied to the videos to extract the continuous body movements (with 13 key points) of each person. Then the extracted key point locations were constructed into time-series inputs for the proposed LSTM-MTL framework for human intention prediction. Following relevant studies in motion prediction, e.g., (Mao et al. 2019), the observation duration was set to 400ms (i.e., 6 frames) and the prediction duration was set to 400ms (i.e., 6 frames) for short-term prediction. The actions and involved objects were manually annotated on a frame-by-frame basis. The dataset includes a total of 27,314 images (i.e., total video length is around 30 min). Table 1 lists the distribution of different actions and objects in the dataset.

Table 1 Sample size for each class of actions and objects

Action Class	**Sample Size**	**Object Class**	**Sample Size**
Pick	3,935	Stone	15,309
Carry	2,621	Connector	8,913
Assemble	7,309	Stone+connector	3,092
Adjust	8,809	*Note: "Stone+connector" is a special object type only in task 3, where connectors are integrated with stones.*	
Inspect	2,123		
Release	2,517		

The dataset was split into 80% for training and 20% for testing. In the training process, 5-fold cross-validation was adopted to select the best configuration of parameters, including batch size, hidden unit, number of epochs, dropout rate, and learning rate. The average performance of the 5 sub-models was treated as the performance under a specific configuration and used to select the best configuration. Once the configuration was selected, the corresponding model was retrained using the entire training set and tested on the testing set to evaluate the performance of the proposed framework. As a result, the LSTM-MTL network was trained on mini-batches, with a batch size of 60, the hidden unit set to 50, and numbers of epochs at 2500. The dropout rate was selected as 0.2. Adam was used as the optimization algorithm with an initial learning rate of 0.0001. The weights of loss functions for the three tasks were set as 1:1:5.

RESULTS

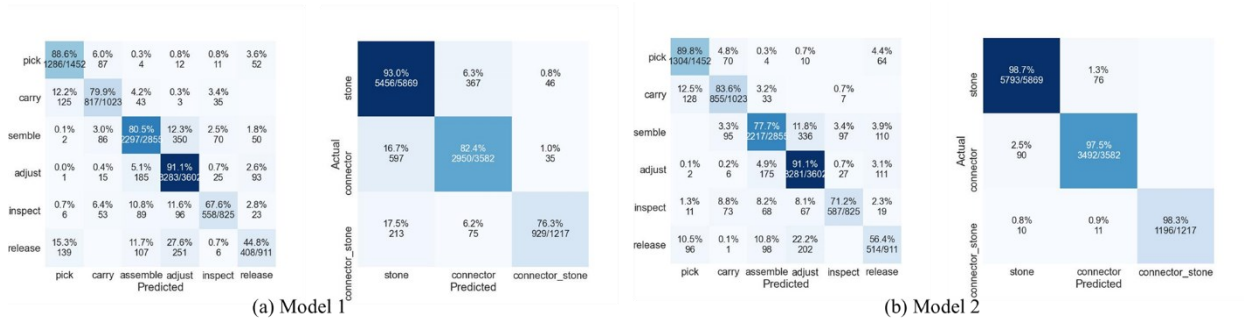
The performance of classification tasks, i.e., action and object prediction, was primarily evaluated using overall accuracy, which measures the percentage of images being correctly classified for all

classes of actions and objects, respectively. The accuracy was computed as $Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \times 100\%$, where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative. The performance of the regression task, i.e., motion prediction, was evaluated using two metrics (Yuan and Kitani 2020): 1) average displacement error (ADE), measuring average L_2 distance between the predicted and ground-truth joint positions over all predicted time steps, computed as $\frac{1}{N \times T}\sum_{i=1}^{N}\sum_{t=0}^{t=T}\left\|\hat{\boldsymbol{y}}_t^i - \boldsymbol{y}_t^i\right\|$. 2) Final displacement error (FDE), measuring L_2 distance between final ground-truth joint positions and predicted joint positions, computed as $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{y}}_T^i - \boldsymbol{y}_T^i\right\|$, where N is sample size, $\hat{\boldsymbol{y}}_t^i$ is predicted pose of i^{th} data at time t, $\boldsymbol{y}_t^i$ is ground truth pose of i^{th} data at time t, and T is prediction duration.

To demonstrate the advantage of incorporating contextual information, two models were compared, where Model 1 indicates the baseline model that does not consider contextual information, and Model 2 incorporates types of involved objects in the observation period in the input features as contextual information. Table 2 lists the results. It can be shown that both models can predict actions and objects at a satisfactory accuracy (over 80%). Specifically, incorporating contextual information leads to a 11.36% improvement in object prediction because the tasks involve pre-defined assembly order, and the objects needed are heavily depending on the current work status. In contrast, the accuracy of action prediction does not differ significantly between two models since it is primarily inferred from human body poses.

Table. 2 Quantitative results of multi-level intention prediction

Model	Task 1 – Action prediction	Task 2 – Object prediction	Task 3 – Body motion prediction	
	Accuracy	Accuracy	ADE (pixel)	FDE (pixel)
Model 1 (baseline)	0.81	0.88	8.78	10.61
Model 2 (context-aware)	0.82	0.98	8.71	10.56



(a) Model 1 (b) Model 2

Figure 6. Confusion matrix for action and object prediction

The classification performance of individual classes in these two tasks is further illustrated using confusion matrix (see Figure 6). In action prediction, "inspect" and "release" exhibit relatively low accuracy, which is potentially due to the smaller sample size. However, incorporating contextual information significantly increases the performance in these two classes. Furthermore, most misclassifications occurred in similar actions that may happen alternatively during a short period, such as "assemble" and "adjust", "release" and "adjust", which may require additional cues to better differentiate them. In terms of body motion prediction, Model 2 shows a slight improvement compared to Model 1. As body pose may be affected by the involved objects

(e.g., different weights and shapes), prior knowledge of such context could lead to a better inference of future motion. A sample prediction result is visualized in Figure 7.

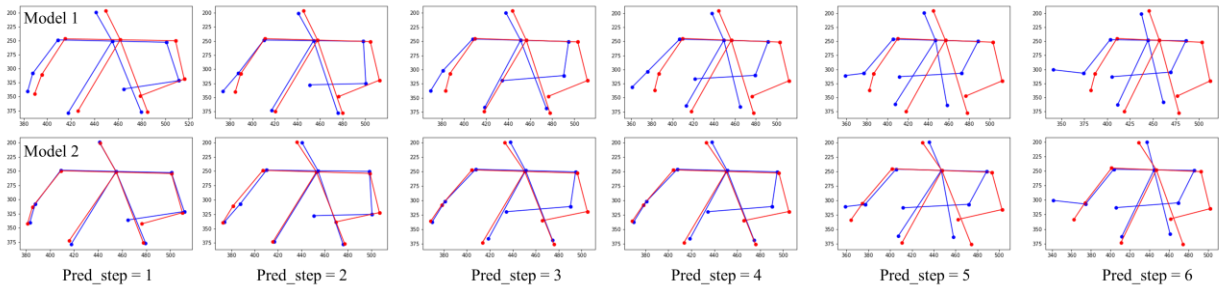


Figure 7. Demonstration of predicted body movements over six prediction steps (top row shows Model 1 result, bottom row shows Model 2 result; blue lines indicate ground truth poses, and red lines indicate predicted poses).

CONCLUSION

This study proposes an integrated LSTM-based MTL framework to predict multi-level human intention in assembly tasks, including high-level actions and objects, and low-level body movements, from observed body movements and associated assembly components extracted from videos. The proposed method is validated using video data from 6 assembly tasks conducted by 9 people. The results show that the proposed method achieves an accuracy of 82% and 98% for action and object prediction, respectively, and an ADE of 8.71 pixels and an FDE of 10.56 pixels in pose prediction. Specifically, the incorporation of work context, represented as the type of objects involved in current progress, significantly improves the accuracy of object prediction by 11.36%, with the performance of other two tasks increasing slightly. The proposed method is expected to enhance the intelligence of collaborative robots by enabling robots to predict "what is needed", "when is needed", and "where to pass the objects" in HRC assembly tasks, and thus planning adaptively to enhance smooth HRC.

There remain some limitations. First, in the proposed framework, 2D pose was estimated from monocular camera. Our ongoing study uses RGB-D camera to capture body motion, and depth information will be recovered for 3D pose prediction. Second, current object information was manually annotated. In future studies, the proposed framework will be integrated with an object detection module to automatically estimate current object information and achieve fully automatic prediction. Third, in this study, human-human interaction was used to collect human intention during assembly tasks, with the premise that HRC could learn from conventions in human-human interactions. In our ongoing study, further experiments were conducted for both HRC and human-human collaboration, and the proposed framework will be trained and tested using data from both scenarios to better understand human behavior in HRC.

ACKNOLWEDGMENTS

This research was funded by the U.S. National Science Foundation (NSF) via Grants 2138514 and 2222670. The authors gratefully acknowledge NSF's supports. Any opinions, findings, recommendations, and conclusions in this paper are those of the authors, and do not necessarily reflect the views of NSF and The University of Texas at San Antonio.

REFERENCES

Cai, J., Li, X., Liang, X., Wei, W., and Li, S. (2022). "Construction Worker Ergonomic Assessment via LSTM-Based Multi-Task Learning Framework." *2022 Construction Research Congress*, 215–224.

Cai, J., Yang, L., Zhang, Y., Li, S., and Cai, H. (2021). "Multitask Learning Method for Detecting the Visual Focus of Attention of Construction Workers." *Journal of Construction Engineering and Management*, American Society of Civil Engineers, 147(7), 4021063.

Cardno, C. A. (2018). "Robotic Rebar-Tying System Uses Artificial Intelligence." *Civil Engineering Magazine Archive*, 88(1), 38–39.

Kim, S., Peavy, M., Huang, P. C., and Kim, K. (2021). "Development of BIM-integrated construction robot task planning and simulation system." *Automation in Construction*, 127.

Krämer, N. C., Von Der Pütten, A., and Eimler, S. (2012). "Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction." *Studies in Computational Intelligence*, 396, 215–240.

Liu, H., and Wang, L. (2017). "Human motion prediction for human-robot collaboration." *Journal of Manufacturing Systems*, 44, 287–294.

Liu, Z., Liu, Q., Xu, W., Liu, Z., Zhou, Z., and Chen, J. (2019). "Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing." *Procedia CIRP*, 83, 272–278.

Luo, R. C., and Mai, L. (2019). "Human Intention Inference and On-Line Human Hand Motion Prediction for Human-Robot Collaboration." *IEEE International Conference on Intelligent Robots and Systems*, 5958–5964.

Madsen, A. J. (2019). "The SAM100: Analyzing Labor Productivity." <https://digitalcommons.calpoly.edu/cmsp/243/> (Jul. 5, 2022).

Mao, W., Liu, M., Salzmann, M., and Li, H. (2019). "Learning trajectory dependencies for human motion prediction." *Proceedings of the IEEE International Conference on Computer Vision*, 9488–9496.

Wang, W., Li, R., Chen, Y., Sun, Y., and Jia, Y. (2022). "Predicting Human Intentions in Human-Robot Hand-Over Tasks Through Multimodal Learning." *IEEE Transactions on Automation Science and Engineering*, 19(3), 2339–2353.

Wibranek, B., and Tessmann, O. (2019). "Digital rubble compression-only structures with irregular rock and 3D printed connectors." *IASS Symposium 2019 - 60th Anniversary Symposium of the International Association for Shell and Spatial Structures; Structural Membranes 2019 - 9th International Conference on Textile Composites and Inflatable Structures, FORM and FORCE*, 2488–2495.

Xiu, Y., Li, J., Wang, H., Fang, Y., and Lu, C. (2019). "Pose flow: Efficient online pose tracking." *British Machine Vision Conference 2018, BMVC 2018*, BMVA Press.

Yuan, Y., and Kitani, K. (2020). "DLow: Diversifying Latent Flows for Diverse Human Motion Prediction." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 346–364.

Zhou, T., Wang, Y., Zhu, Q., and Du, J. (2022). "Human hand motion prediction based on feature grouping and deep learning: Pipe skid maintenance example." *Automation in Construction*, 138.