Worksite Object Characterization for Automatically Updating Building Information Models

Max Ferguson[1], Seongwoon Jeong[2] and Kincho H. Law, PH.D., DIST.M.ASCE[3]

[1]Engineering Informatics Group, Department of Civil and Environmental Engineering, Stanford University, 473 Via Ortega, Stanford, CA 94305; e-mail: maxferg@stanford.edu
[2]Engineering Informatics Group, Department of Civil and Environmental Engineering, Stanford University, 473 Via Ortega, Stanford, CA 94305; e-mail: swjeong3@stanford.edu
[3]Engineering Informatics Group, Department of Civil and Environmental Engineering, Stanford University, 473 Via Ortega, Stanford, CA 94305; e-mail: law@stanford.edu

ABSTRACT

Automated data capture systems could significantly improve the efficiency and productivity of the architecture, engineering, construction and facility management (AEC/FM) industry. However, automatically collecting spatiotemporal information in an unstructured environment such as a construction site or a work place remains a time consuming and challenging task. This paper presents a new approach to automated data capture and processing, referred to as object characterization. In object characterization, the goal is to identify common objects in a scene and extract rich semantic information about those objects. A novel 2D-3D object detection algorithm is designed for detection and characterization of common worksite objects. The proposed system has applications in automated surveying and data collection, especially in applications which leverage unmanned aerial vehicles or mobile robots. To demonstrate this utility, the proposed system is deployed on a mobile robot and used to detect newly placed objects in a worksite environment.

INTRODUCTION

The growing prevalence of automated data capture systems and semi-autonomous robots is likely to have a positive impact on efficiency and productivity in the construction and facility management sectors. Recent advances in Lidar, depth-sensing, visual odometry, and related technologies (Engelhard et al. 2011; Kerl et al. 2013) have made it possible to obtain a static representation of a large space with relative ease. In most cases, this static representation exists as a point cloud or polygon mesh. Research efforts in computer vision have focused on the segmentation of these point-clouds into a number of preset classes (Armeni et al. 2016; Tchapmi et al. 2017). However, many applications in construction and facility management require much richer information about the objects. Ultimately, a mobile robot or unmanned aerial vehicle (UAV) should be able to scan a worksite and return a wealth of information about the objects in that environment. As a step toward this goal, this paper focusses on developing a computer vision system to infer semantic-rich information about worksite objects from RGB-D images.

Previous systems in civil engineering have been focused on the tasks of image classification (Gao and Mosalam 2018), image segmentation (Feng et al. 2017), and object detection (Ferguson and Law 2019). This paper introduces a much richer computer vision task called object

Figure 1. Object detection and characterization for three worksite objects, using the proposed object characterization framework and algorithm.

characterization. Object characterization is defined as the process of inferring physical attributes about one or more objects in an image or 3D scan. Formally, each object is assigned an object class, as in object detection. Each object class is associated with a set of attributes. Some attributes may be shared across all the object classes, such as the object size, material and color. Other attributes may be class-specific, such as the fullness of a recycle bin. The goal is to correctly predict the object class and all of the applicable object attributes. Some examples of object characterization are provided in Figure 1.

In a physical engineering environment, object characterization offers many benefits over related tasks in automated data capture:

- **Increased semantic information:** The ability to predict object attributes in addition to object class makes object characterization much more valuable than object detection alone.
- **Structure through ontology:** As ontology is used to represent prior knowledge about objects and their attributes, a predictive model for object characterization can be used to populate the ontology using images or Lidar scans from the worksite.

This paper introduces a methodology that uses object characterization to automatically populate a building information model with semantic-rich information inferred from RGB-D images. The following summarize the main contributions of this paper:

- A generic framework for object characterization is proposed.
- A new 2D-3D object detection algorithm is developed for the detection and characterization of worksite objects.
- A dataset containing RGB-D images of common worksite objects is collected and annotated with a rich set of attribute labels which can be used by other researchers.

The remainder of the paper is organized as follows: An overview of related works is provided in the next section. A detailed description of the proposed object characterization system is then presented. For validation, experiments are conducted to demonstrate how the proposed algorithm can be used to automatically update a building information model. The paper is concluded with a brief summary and discussion.

RELATED WORK

To further enhance current practice on construction automation, research efforts have been directed towards the problem of automatically extracting semantic information from 2D images (Zhu et al. 2016) and 3D point clouds (Rusu et al. 2008). However, engineering applications generally require much richer information from data capture systems than that provided by semantic

segmentation. The Structural ImageNet project is one example that attempts to provide one such dataset where objects are labelled with semantic attributes like material type and damage state (Gao and Mosalam 2018). However, Structural ImageNet is posed as a classification task, where individual images are assigned a class based on their content. Object detection and segmentation tasks are much more suited to engineering applications classification tasks, as they identify the position of relevant objects within each image. As far as we are aware, there is yet to be a generic framework for the prediction of object attributes using RGB-D or Lidar images.

Recently, there has been some interest in using unmanned aerial vehicles (UAVs) or unmanned ground vehicles (UGVs) to automatically gather information about the building environment (Hamledari et al. 2017b; Park et al. 2016). However, the computer vision systems so far tend to be incomplete, often lacking the ability to accurately classify or localize objects of interest (Asadi and Han 2018; Hamledari et al. 2017a; McCabe et al. 2017). There has been some interest in updating IFC models using semantic-rich input data (Hamledari et al. 2017b). However, most previous approaches are impractical as they propose 2D image processing algorithms for the purpose of updating 3D models.

The notion of automatically creating geometric environment maps has been explored extensively by the robotics community. Perhaps the most noteworthy technique is Simultaneous Localization and Mapping (SLAM), which aims to construct a map of an unknown environment while simultaneously keeping track of an agent's location within it (Thrun and Leonard 2008). Related tasks in robotics include visual-SLAM and structure from motion (SfM), which generate a static geometric representation of the environment using a consecutive series of RGB images (Engelhard et al. 2011). While effective for reconstructing scene geometry, these techniques do not directly return semantic information about the environment.

WORKSITE OBJECT CHARACTERIZATION

This section focuses on the development of an algorithm for the detection and characterization of worksite objects, using RGB-D images from a mobile robot. The proposed system, referred to as Extended Object R-CNN, first detects objects of interest using a deep convolutional neural network (CNN). Features from the CNN are then used to predict the remaining object attributes. The output of the system is a list of detected objects populated with attribute information for each object.

Object Ontology: The object characterization task is based on an ontology which defines object classes and their attributes. Object classes could include entries like *recycle bin*, *hard hat* or *gas container*. Each class is associated with one or more attributes. For example, attributes for the recycle bin class could be *color, material, size* or *fullness.* Attributes can be numerical, categorical, or Boolean. Numerical attributes, such as color are represented as vectors. Categorical attributes can take on a single value from a predefined a set of potential values. Boolean attributes can take on a true or false value, as in the case of *fullness.* An example ontology for the recycle bin class is provided in Figure 2.

Object R-CNN: An algorithm is developed to simultaneously address the tasks of object detection and object characterization. The proposed algorithm builds on the Object R-CNN framework, that has been described in an earlier work (Ferguson and Law 2019). Object R-CNN is a collection of CNNs designed to detect and localize worksite objects in 3D space. In Object R-CNN, each image is first passed through a CNN to develop a featurized representation of the image. A feature pyramid network (FPN) with an InceptionResNetV2 backbone is used as the feature extractor (Szegedy et

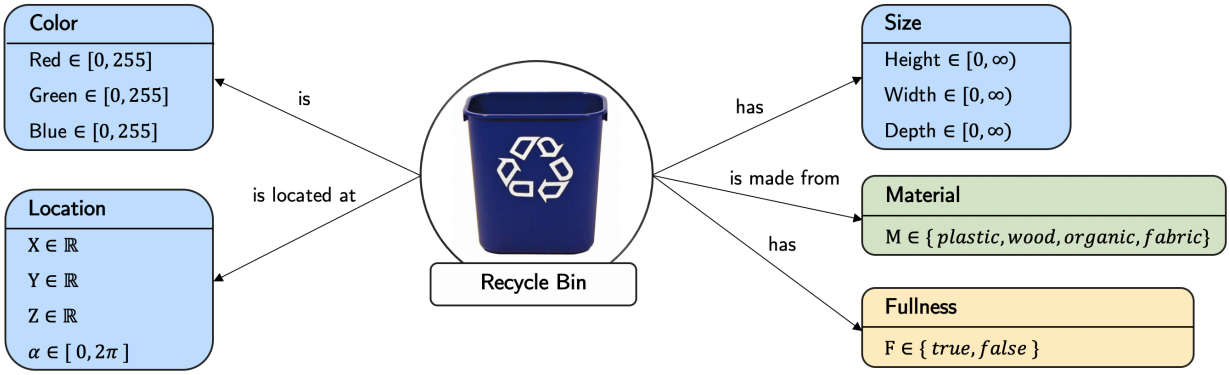

Figure 2. Example ontology for the recycle bin class.

al. 2017). In the next stage, a region proposal network (RPN) is used to generate a set of regions of interest (RoIs) that are likely to contain objects. Each RoI is assigned a bounding box, class and segmentation mask by a second CNN. Finally, each detection is projected into 3D space to form a 3D region proposal. Features are cropped from the RGB-D point cloud using this region proposal and fed into a 3D CNN. The output from the 3D CNN is used along with the 2D region-based features to generate a refined estimate of the size and position of the object in 3D space.

Extended Object R-CNN: Figure 3 shows the Extended Object R-CNN architecture that is designed for the prediction of arbitrary object attributes. It is hypothesized that concatenating features from the 2D image and the 3D voxel representation would provide a rich mathematical representation of each object that is sufficient for the task of object characterization. In Extended Object R-CNN, the fully connected neural network used to predict size and position is replaced with a similar, but larger neural network that has one head for each possible object attribute, allowing it to simultaneously predict every attribute for a given object. The benefit of this approach is that the data associated with multiple different object classes can be used to train a predictor for a common attribute, such as object color. On the contrary, the neural network must predict values for the set of all attributes in the ontology, not just those applicable to the object class.

Loss Functions: Neural networks are trained by minimizing a loss function with respect to the neural network parameters. The output of a neural network branch is denoted as $f(a; I, \theta)$, where a is an anchor location in the image, I is the image, and θ is the neural network parameters. An encoding ϕ is used to convert object attributes to a vector form. A loss function captures the difference between the ground truth encoding and the output of the neural network. To apply the Extended Object R-CNN framework to a new attribute, it is necessary to choose an encoding and loss function for that attribute.

To illustrate, the following example shows the encoding and loss function for a color attribute. The dominant color of an object is defined as the color which covers the greatest portion of the object. The color is represented as a three-element vector with components red, $r \in [0,255]$, green $g \in [0,255]$ and blue $b \in [0,255]$. The color encoding ϕ_{color} is chosen to map each component onto the domain $[0,1]$:

$$\phi_{color}(r, g, b; a) = \frac{1}{255}[r, g, b]^T \quad (1)$$

where a is the anchor box (position of the object in the image). The loss function $L_{color}(a; I, \theta)$

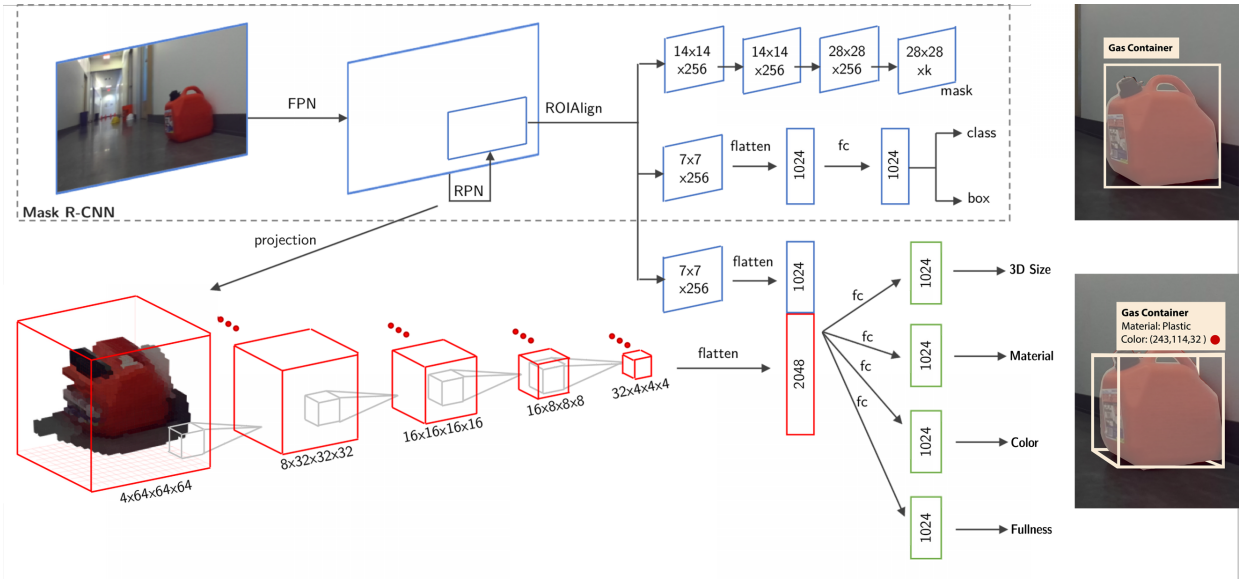

Figure 3. Extended Object R-CNN builds on Object R-CNN by adding an additional network head for each attribute defined in the object ontology. The numbers in figure indicate the dimensions of intermediate representations (tensors). Arrows denote convolutional layers unless labelled otherwise. The term "fc" represents a fully connected neural network layer.

measures the difference between the encoded color and the neural network output and can be expressed using a 2-norm as:

$$L_{color}(a; I, \theta) = \|\phi_{color}(r, g, b; a) - f_{color}(a; I, \theta)\|_2^2 \quad (2)$$

In general, a simple linear transformation can be used to encode numerical attributes, and a one-hot encoding can be used to encode categorical attributes (Goodfellow et al. 2016). The quadratic loss function can be used for numerical attributes and the cross-entropy loss function can be used for categorical and Boolean attributes. The original Object R-CNN framework defines loss functions for the predicted class L_{cls}, bounding box L_{bbox}, and segmentation mask L_{mask} (Ferguson and Law 2019). In Extended Object R-CNN, the total loss for each anchor a is defined as the sum of L_{cls}, L_{bbox}, L_{mask} and the loss from each of the attributes. The loss for each RGB-D image is defined as the average loss across positive anchors.

EXPERIMENTS

The proposed Extended Object R-CNN system is trained and tested on a set of RGB-D images collected with a purpose-built SLAM-enabled mobile robot.

Dataset: Objects in the RGB-D images are annotated with class labels and 2D segmentation masks. A total of 7423 objects are annotated across 1214 RGB-D images to form a training dataset. A further 1214 objects across 255 images are annotated to form the test set. Each object is annotated with the additional attributes: *3D bounding box*, *3D position*, *dominant material*, and *dominant color*. The dominant material of each object is chosen from the set $\{plastic, fabric, wood, organic\}$, where the *organic* class is used to described plant material. Identifying the true color of an object from a photo is challenging, due to the effects of illumination, shadow, and camera properties (Zaidi and Bostic 2008). The ground truth color is assigned by matching the color from a color palette app on a mobile phone to the real object, under bright white light. The recycle bin and bucket objects

are also annotated with the binary attribute, *fullness*. Fullness is deemed true if any object can be seen protruding from the container, otherwise it is labelled as false.

Implementation Details: Extended Object R-CNN is trained on the aforementioned dataset. The neural network weights for Mask R-CNN are initialized using pretrained weights generated using the Microsoft Common Objects in Context (COCO) dataset (Lin et al. 2014). The remaining weights are initialized randomly using the Xavier initialization method (Glorot and Bengio 2010). The model is trained by minimizing the total loss across all images in the training dataset, using stochastic gradient descent with momentum.

Training a neural network with a large number of network heads can be unstable. For example, large errors in the color prediction may propagate back into the convolutional layers and cause the CNN weights to change substantially. To overcome these instability issues, the following training scheme is adopted: The neural network is first trained to predict 2D bounding boxes and segmentation masks, as described in (He et al. 2017). Next, all of the neural network weights except the attribute head weights are frozen. The network is then trained to predict object attributes for 20 epochs. Finally, all of the weights are unfrozen, the learning rate is dropped by a factor of 10, and the network is trained for a further 20 epochs.

Evaluation Metric and Results: The system is evaluated on the test dataset. The 11-point average precision (AP) metric is used as a metric for quantifying the 3D bounding box accuracy (Manning et al. 2008). A prediction is considered correct if the IoU_{vol} of the predicted bounding box and the ground truth is larger than 0.5. The prediction of other attributes like color and material is only evaluated if the 3D bounding box is predicted correctly. A color prediction is considered correct if all of the RGB values are within 10% of the ground truth. The results are presented in Table 1. Finally, the object detection and characterization algorithm is combined with the model updating system described in (Ferguson and Law 2019) to automatically update a building information model in real time with rich information about newly spawned objects. The end-to-end system for automated data capture, object characterization and model updating is summarized in Figure 4.

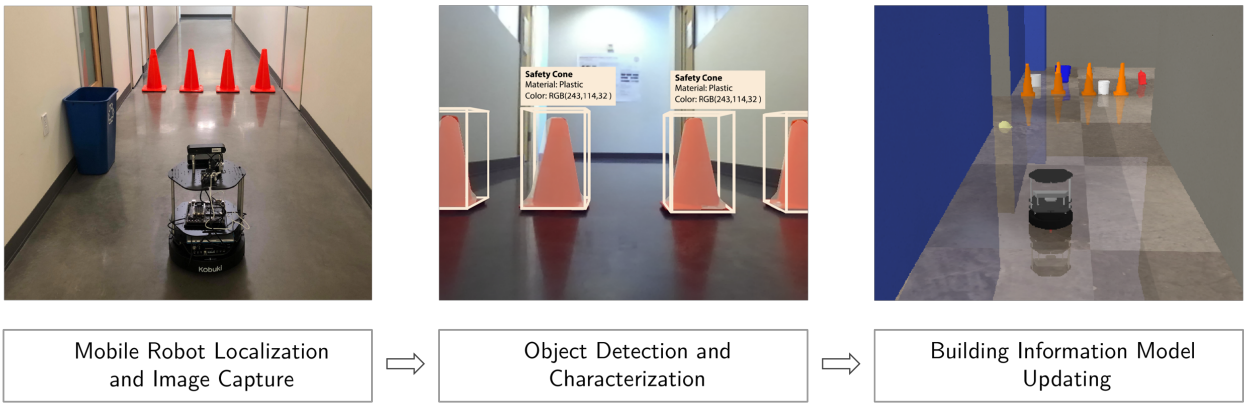

Figure 4. Automatically updating a building information model using RGB-D frames collected with a mobile robot. Each RGB-D frame is processed using the proposed object detection and characterization algorithm, to extract semantic-rich information. This information is then used to dynamically update a building information model.

Table 1. Accuracy of Extended Object R-CNN network on bounding box prediction and attribute prediction.

Class	*Box Prediction (AP)*	*Color (%)*	*Material (%)*	*Fullness (%)*
Safety Cone	0.95	83.1	94.2	-
Hard Hat	0.83	96.2	99.1	-
Bucket	0.84	94.1	95.7	76.3
Recycle Bin	0.86	90.4	90.4	85.3
Safety Gloves	0.79	78.2	81.9	-
Gas Container	0.87	84.5	83.7	-
Chair	0.90	85.3	94.3	-
Plant	0.88	70.3	93.1	-
Overall	0.87	87.9	91.8	81.3

SUMMARY AND DISCUSSION

This paper presented a framework and computer vision system for detection and characterization of worksite objects, using RGB-D images from a mobile robot. Specifically, the paper introduced Extended Object R-CNN, a new computer vision system for worksite object detection and characterization. Experimental results have so far demonstrated that Extended Object R-CNN could detect worksite objects in RGB-D images, and then populate an ontology for each object. The proposed framework supports both categorical and numerical attributes making it suitable for a range of automated data capture tasks.

The proposed framework and algorithm enables semantic-rich information to be automatically extracted from RGB-D images. The framework is sufficiently flexible to make it useful in a large number of civil engineering applications. The 3D position of each object is predicted, allowing detected objects to be added directly to a building information model. Different object types can be added to the ontology without changing the underlying algorithms. New attributes can be added by simply specifying an encoding function and a loss function. For most categorical variables, good results are observed using a one-hot attribute encoding coupled with the cross-entropy loss function. Similarly, with continuous numerical variables, the quadratic loss function works well with any reasonable choice of encoding.

The proposed Extended Object R-CNN algorithm is not without limitations. The algorithm is computationally expensive to train and evaluate, requiring several days to train on a powerful desktop computer with a single GPU. The evaluation rate was limited to approximately 0.9 frames/second on the same hardware. Furthermore, the algorithm may not learn effectively when training on datasets where certain object attributes are seen very infrequently.

ACKNOWLEDGEMENT

This research is partially supported by the Center for Integrated Facility Engineering at Stanford University. The first author is also supported by the John A. Blume fellowship.

REFERENCES

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. (2016). "3D semantic parsing of large-scale indoor spaces." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 1534–1543.

Asadi, K., and Han, K. (2018). "Real-Time image-to-BIM registration using perspective alignment for

automated construction monitoring." *Construction Research Congress*.
Engelhard, N., Endres, F., Hess, J., Sturm, J., and Burgard, W. (2011). "Real-time 3D visual SLAM with a hand-held RGB-D camera." Vasteras, Sweden, 1–15.
Feng, C., Liu, M.-Y., Kao, C.-C., and Lee, T.-Y. (2017). "Deep active learning for civil infrastructure defect detection and classification." *ASCE International Workshop on Computing in Civil Engineering*, Seattle, WA, United States, 298–306.
Ferguson, M., and Law, K. (2019). "A 2D-3D Object Detection System for Updating Building Information Models with Mobile Robots." *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa Village, HI, USA.
Gao, Y., and Mosalam, K. M. (2018). "Deep transfer learning for image-based structural damage recognition." *Computer-Aided Civil and Infrastructure Engineering*, 33(9), 748–768.
Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks." *Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy, 249–256.
Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*. MIT press Cambridge.
Hamledari, H., McCabe, B., and Davari, S. (2017a). "Automated computer vision-based detection of components of under-construction indoor partitions." *Automation in Construction*, 74, 78–94.
Hamledari, H., Rezazadeh Azar, E., and McCabe, B. (2017b). "IFC-based development of as-built and as-is BIMs using construction and facility inspection data: Site-to-BIM data transfer automation." *Journal of Computing in Civil Engineering*, 32(2).
He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN." *IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, 2980–2988.
Kerl, C., Sturm, J., and Cremers, D. (2013). "Dense visual SLAM for RGB-D cameras." *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo Big Sight, Tokyo, Japan, 2100–2106.
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). "Microsoft COCO: Common objects in context." *European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 740–755.
Manning, C., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
McCabe, B., Hamledari, H., Shahi, A., Zangeneh, P., and Azar, E. R. (2017). "Roles, benefits, and challenges of using UAVs for indoor smart construction applications." *ASCE International Workshop on Computing in Civil Engineering*, 349–357.
Park, J., Cho, Y. K., and Martinez, D. (2016). "A BIM and UWB integrated mobile robot navigation system for indoor position tracking applications." *Journal of Construction Engineering and Management*, 6(2), 30–39.
Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M. (2008). "Towards 3D point cloud based object maps for household environments." *Robotics and Autonomous Systems*, 56(11), 927–941.
Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning." *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, San Francisco, California, USA, 12.
Tchapmi, L. P., Choy, C. B., Armeni, I., Gwak, J., and Savarese, S. (2017). "SEGCloud: Semantic segmentation of 3D point clouds." *arXiv preprint arXiv:1710.07563*.
Thrun, S., and Leonard, J. J. (2008). "Simultaneous localization and mapping." *Springer Handbook of Robotics*, Springer, 871–889.
Zaidi, Q., and Bostic, M. (2008). "Color strategies for object identification." *Vision research*, 48(26), 2673–2681.
Zhu, Z., Park, M.-W., Koch, C., Soltani, M., Hammad, A., and Davari, K. (2016). "Predicting movements of onsite workers and mobile equipment for enhancing construction site safety." *Automation in Construction*, 68, 95–101.