A Timely Object Recognition Method for Construction Using the Mask R-CNN Architecture

D. Shamsollahi[a], O. Moselhi[b] and K. Khorasani[c]

[a]Department of Building, Civil and Environmental Engineering, Concordia University, Canada
[b]Centre for Innovation in Construction and Infrastructure Engineering and Management (CICIEM), Department of Building, Civil and Environmental Engineering, Concordia University, Canada
[c]Department of Electrical and Computer Engineering, Concordia University, Canada

E-mail: D_shams@encs.concordia.ca, Moselhi@encs.concordia.ca, kash@ece.concordia.ca

Abstract –

Efficient progress monitoring and reporting require detailed and accurate reports from construction sites in a timely manner. These reports include important information to assist decision-making through comparison of as-built information to as-planned state. Manual reporting is time-consuming, error-prone, costly and is highly dependent on site personnel expertise. Advances recently made in artificial intelligence, data processing and digital cameras have paved the way for introduction of image-based methods for automated monitoring and progress reporting in the construction industry. Object recognition has achieved significant advances and considerable growth by the introduction of deep learning algorithms such as the Convolutional Neural Networks (CNN). This research proposes a method for automated recognition and segmentation of HVAC ducts utilizing digital images by developing Mask Region-based Convolutional Neural Network (Mask R-CNN) architectures. 3D BIM models are utilized for generating 1,143 synthetic images to train the developed Mask R-CNN model. To enhance the training dataset capability and overcome the overfitting problems, various data augmentation techniques are considered. The developed deep learning-based object recognition method automates monitoring of HVAC ducts installation, making use of generated synthetic images for training the algorithm to overcome the need for large datasets of actual images.

Keywords –

Deep Learning; Convolutional Neural Networks (CNN); Mask R-CNN; Progress Monitoring

1 Introduction

Constant progress monitoring of different activities at jobsites influences the project budget and schedule for reducing cost and delays. It also improves the quality control, documentation, and communication in construction projects [1]. In conventional progress monitoring schemes, the current state of the project is compared with the as-planned state to assist in evaluating the project's performance. It includes tracking, reviewing, and organizing the activities for determining the areas where timely corrective actions are required [2, 3].

However, due to different site activities, monitoring the construction progress is a complicated and challenging task that requires correct information in a timely manner to support project managers in identifying scheduling deviations in an early phase to avoid possible delays. Accurate and in-time construction site data collection, efficient data analysis, and visualization applications in an interpretable format are essential requirements for efficient progress monitoring [2, 3].

Currently, many construction sites are equipped with economical digital cameras that produce a large number of images and videos containing considerable amount of information from the job sites. These images/videos can ultimately benefit the project management system. However, due to various difficulties in data analysis and processing, practical use of this abundant of data is quite challenging. Hence, project managers would apply manual and costly methods for construction activity analysis [4].

Through improvements in deep learning algorithms and advances in device capabilities (processing power, memory storage and high image sensor resolution), computer vision methodologies have gained widespread interest in various construction research areas [5].

Consequently, by increasing efficiencies in extracting information from the captured images and videos, computer vision methods that use deep learning algorithms have been applied for automating construction monitoring purposes such as in progress tracking, productivity analysis, safety assurance, and quality control [6, 7].

In particular, there is an increasing shift towards utilizing deep-learning based object recognition algorithms for automated identification of construction elements from digital images to assist project reporting and updating schedule by accessing to as-built information [4, 7–9].

However, despite advances in different object recognition algorithms, the open dataset of images from construction job sites, including different building elements, is not available to train and validate the algorithms [4]. This research has developed a method to automatically recognize HVAC ducts using Mask R-CNN architecture and evaluate the model by utilizing quantitative performance metrics.

Towards this end, 3D BIM models were utilized to generate and extract synthetic images to train the CNN algorithm. Moreover, image augmentation techniques such as geometric transformations and kernel filters were applied to artificially increase the training dataset size for a stable network training and prevent overfitting. Two experiments were conducted to evaluate data augmentation impact in the ultimate HVAC ducts recognition performance.

2 Related Work

Computer vision methodologies can facilitate the construction monitoring systems through detecting and tracking material, equipment, and labor in construction job sites [1, 4, 7, 10–14]. In computer vision, the detection and classification of objects in images/videos can be categorized into traditional (feature-based) algorithms and deep learning algorithms [7, 15].

In traditional algorithms, human-engineered features such as edges, corners, and colors are extracted to determine the correct class of objects [12, 15]. This is achieved by utilizing examples of feature descriptors such as Haar-like, the histogram of oriented gradients (HOG), Speeded-Up Robust Features (SURF), color histograms, among others. These feature descriptors mostly are combined with machine learning algorithms that have shallow structures such as the K-Nearest Neighbors and the Support Vector Machine for conducting the classification tasks [7, 15, 16].

A number of research studies have utilized feature descriptors and machine learning algorithms to detect construction resources from digital images [1, 17–19].

Deep Learning (DL) algorithms such as the Convolutional Neural Networks (CNN), have shown promising performance in object detection areas and are widely applied in the construction industry. These methodologies provide more practical solutions through self-learning capability and higher accuracy when compared to the traditional algorithms [12]. Different studies have used deep learning algorithms to detect objects of interest in construction sites. The reference [4] has detected the structural components such as beams and columns by utilizing Deeply Supervised Object Detector (DSOD), [9] has applied Mask R-CNN, a deep convolutional neural network to detect Walls, Doors, and Lifts from images for creating an as-built model.

Finally, the refence [10] has developed a framework including Convolutional Neural Networks for detecting the existing building objects in the jobsite, and then the extracted objects were superimposed on the as-planned model through BIM and Virtual Reality to evaluate the progress state. Despite the significant performance of deep learning algorithms to detect construction components with high accuracy, there is still currently lack of large, labeled image datasets from job sites including different classes in the construction industry [20]. Hence, the application of synthetic images for training deep learning algorithms for performing object detection purposes has received a significant amount of interest to overcome the above issues [21, 22].

3 Developed Method

Figure 1 provides an overview of the developed method. It consists of three main modules, namely: “Synthetic Image Generation and Data Labeling”, “Mask R-CNN Training”, “Testing and Evaluation”. Each of these modules and components is described in the following sections.

3.1 Synthetic Image Generation and Data Labeling

Similar to the recent research conducted in [23], 3D BIM models are utilized to extract synthetic images for overcoming absence of construction elements real image datasets for training the deep learning algorithm. In our work, the BIM models are taken from an online open-source website [24]. The HVAC ducts and other building elements properties such as shape, material and size are defined in Autodesk Revit 2019 as a BIM software.

Different viewpoints that consider various occlusions and illuminations are selected and rendered using Enscape v2.8, which is a real-time rendering Plugin installed in the Revit software. A total of 1,143 synthetic images are generated and used for network training. The dataset consists of one class, namely HVAC duct which is a common and widely used element in building construction. It has 1,887 duct instances captured in

1,143 images. The training set distribution shows that from 1,143 images, 56% of images contain only one HVAC duct in each image, 32% have two ducts, 9% three ducts, 2% four ducts, and 1% five ducts. 172 images are randomly selected for testing and validation purposes. The test set data follows nearly the same distribution of the training; 54% of images having one HVAC duct, 44% having two ducts, and 2% having three ducts. The images are manually annotated using the VGG Image Annotator (VIA) web tool to specify the HVAC ducts regions and locations in images through polygon shapes. In each image, all the pixels that have not been assigned to HVAC duct class are categorized as background.

The annotation files are downloaded as JSON format containing polygons' coordinates of all the images for further model training.

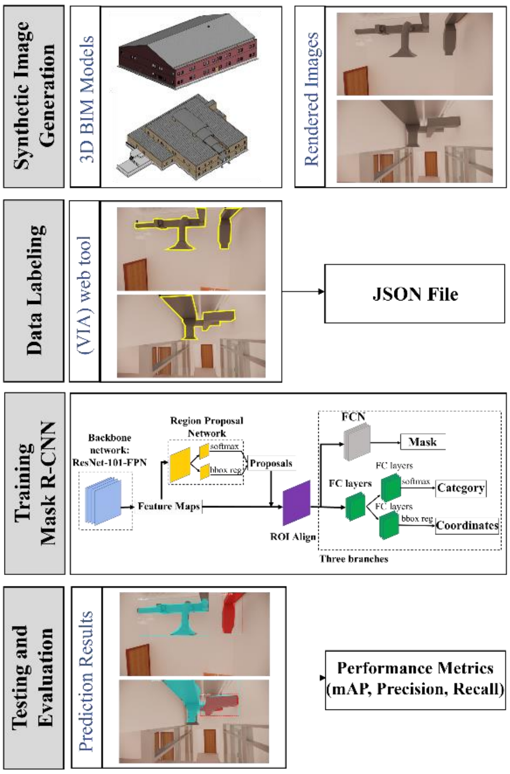

Figure 1. Overview of the proposed method

3.2 Instance Segmentation using Mask R-CNN

For our research we have employed the Mask R-CNN, an instance segmentation technique which is an extension of Faster R-CNN. In this method as compared to Faster R-CNN, a mask prediction branch is added in parallel with the existing classification and localization of candidate objects in the images.

The detailed Mask R-CNN architecture is depicted in Figure 2. In the Mask R-CNN, convolutional backbone network including ResNet-101 and Feature Pyramid Network (FPN) extract feature maps from an image. Next, the feature maps are fed into Region Proposal Network (RPN) to propose the Regions of Interest (RoIs). Also, the Mask R-CNN is utilizing a quantization-free layer, called RoI Align for extracting predefined size feature maps from each RoI.

In the head network, fully connected layers perform object classification and bounding box regression in each RoI in parallel with a branch for predicting masks (by classifying each pixel into a predefined object class) using a fully convolutional network (FCN). Equation (1) is the multi-task loss function on each RoI referring to the sum of classification loss ($Lcls$), the bounding-box loss ($Lbox$), and the mask loss ($Lmask$). Specifications of $Lcls$ and $Lbox$ which have the same loss functions that are utilized in Faster R-CNN and demonstrate classification and detection error are described in [25] and details of $Lmask$ are provided in [26], where

$$L = Lcls + Lbox + Lmask \qquad (1)$$

3.2.1 Training the model

Training the Mask R-CNN is based on the Matterport's implementation [27] using the open-source libraries Keras and Tensorflow. Feature Pyramid Network (FPN) and ResNet101 are applied as a backbone network and rather than training the model from scratch, it is initialized by utilizing pre-trained weights on the MS COCO dataset. After testing different epochs for training the Mask R-CNN, the best results are achieved with 90 Epochs and the batch size of 2, the weight decay of 0.0001, and the learning rate of 0.001. To minimize the overfitting problem and to improve the generalization of the model, different sets of image augmentation techniques such as the Horizontal Flip, the Vertical Flip, Rotation, the Gaussian Blur and Brightness are investigated to create modified copies of the existing data. Details of this investigation are provided in Table 1.

Table 1 The parameters of the augmentation techniques

Data Augmentation Technique	Parameters
Flip	Horizontal & Vertical
Rotation	One of ϴ=90°, 180°,270°
Brightness (Multiply)	(adding value) (0.8,1.5)
Image smoothing (Gaussian blur)	(σ value of Gaussian kernel) (0.0,5.0)

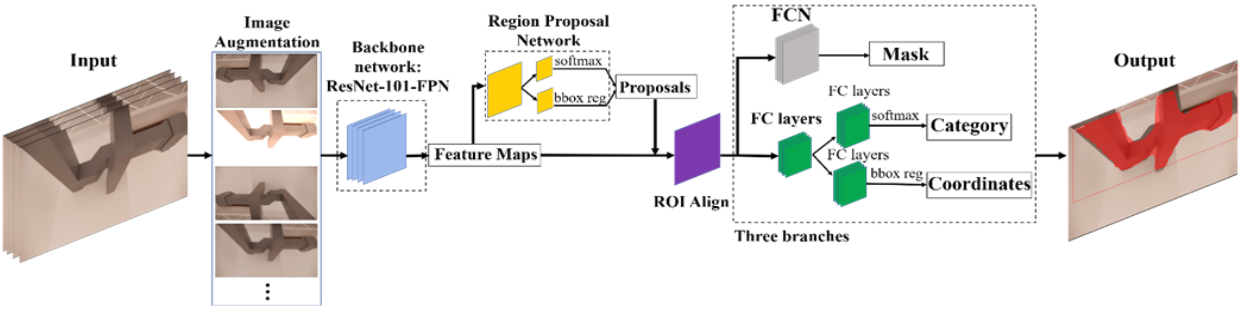

Figure 2. Mask R-CNN network architecture

3.3 Testing and Evaluation

The Mask R-CNN performance can be measured based on the test dataset. Precision and Recall are the selected evaluation metrics. Precision is defined as the ratio of the true predicted samples to the total samples and recall is defined as the ratio of true predicted samples to the total predicted samples, where TP is True Positive, FP is False Positive, and FN is False Negative, that is

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The mean Average Precision (AP) is another commonly utilized metric for evaluation of the CNN object detectors which is defined as the mean of AP over all classes. Based on the Pascal VOC2010–2012 definition, for a pre-defined Intersection over Union value (IoU) as a threshold, AP represents as the area under the precision-recall curve, which is between 0 to 1 and is calculated as follows [28, 29].

$$AP_{all} = \sum_{n} (R_{n+1} - R_N) P_{interp}(R_{n+1}) \tag{4}$$

where the interpolated precision (P_{interp}) at a recall level (R_{n+1}) is equal to the maximum precision that is achieved for any recall level $\tilde{R} \geq R_{n+1}$, that is

$$P_{interp}(R_{n+1}) = \max_{\tilde{R}:\tilde{R} \geq R_{n+1}} P(\tilde{R}) \tag{5}$$

4 Results

In this section, the performance of the Mask R-CNN architecture is evaluated. The model is implemented in the Google Colaboratory (Pro) which is a cloud service based on Jupyter Notebooks with a Tesla P100-PCIE-16GBGPU (accessible up to 24 hours), and the Python3 runtime to overcome the limitations of computer hardware such as disk space for data storage or data processing speed. According to the provided information, the training of the model took 4-5 hours.

To assess effects of augmentation techniques on HVAC duct detection, two experiments are conducted on the training image dataset. In the first experiment, the images are used for training the model with no augmentation technique used (Experiment #1). In the second experiment (Experiment #2), the augmentation techniques that are described in Table 1. are applied for the model regularization and generalization,

The results of the experiment are summarized in Table 2. Moreover, the mAP score for entire images in the Experiment #1 and Experiment #2 is 88.69% and 90.6%, respectively. Figure 3. illustrates the downward trend of the loss function during the training process in the Experiment #2. It also shows the success of the model in preventing overfitting since there is a desired convergence of the training and validation errors. The output images from the HVAC duct detection extracted from the Experiment #2 by using the Mask R-CNN are depicted in Figure 4.

Table 2. HVAC duct detection results using two augmentation experiments.

Training dataset	TP	FP	FN	Precision (%)	Recall (%)
Experiment#1	223	74	32	75.08	87.45
Experiment#2	224	53	31	80.87	87.84

5 Discussion

Due to absence of an open image dataset from building elements in the construction job sites, this research has utilized synthetic images that are generated from 3D BIM models to train a deep learning model for HVAC ducts recognition. Since the dataset is not large, to increase the model performance during the training process, transfer learning based on the COCO dataset, and different data augmentation techniques are considered and applied.

It has been shown in Table 2 that the performance of Experiment #2 with a precision value of 80.87% and a mAP score of 90.6% is better than the Experiment#1 with a precision value of 75.08% and mAP value of 88.69%

and it can be stated that the effect of data augmentation on the model is meaningful and helpful.

According to the results obtained, the developed method provides a robust and accurate tool for recognition of installed HVAC ducts utilizing synthetic images. As such it addresses scarcity of available datasets of real images. It is suggested that performance of the developed method be evaluated using a larger training dataset that includes a mix of both the synthetic and real images and impact of various augmentation techniques be considered for further investigation. Moreover, the developed method can be extended to detect additional building elements such as piping, beams, among others to help progress reporting to be more efficient.

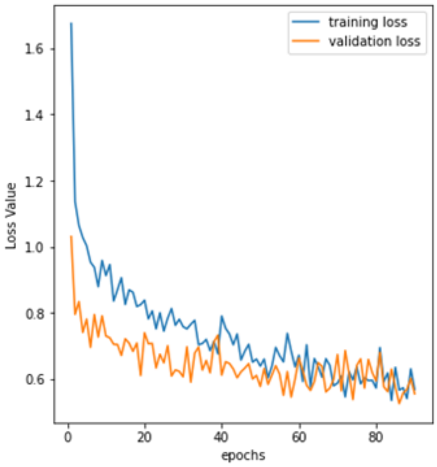

Figure 3. Loss value at each epoch in training and validation sets.

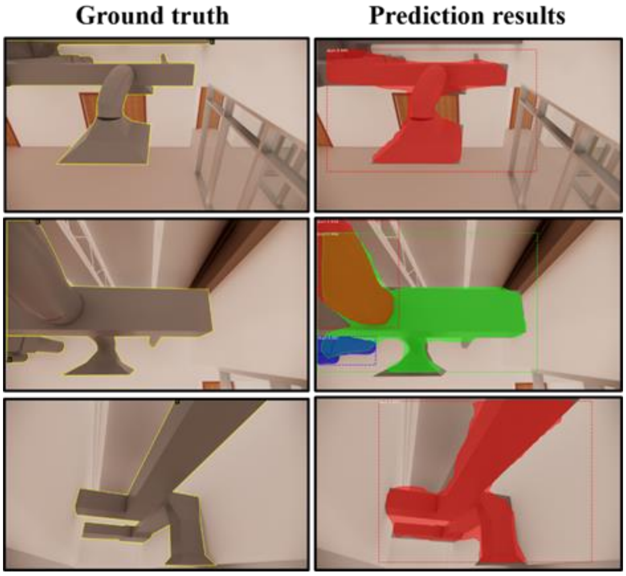

Figure 4. Results of the proposed method

6 Conclusion

This study has utilized a synthetic image dataset that is generated from 3D BIM models to overcome the requirement of having large real-image datasets for training deep learning algorithms. The dataset includes images from various viewpoints, lighting conditions and occlusions to evaluate the robustness of the model. Due to the special appearance of the HVAC ducts, a pixel-wise segmentation approach was selected to increase the accuracy of the detected HVAC ducts spatial locations in images as compared to other algorithms such as the Faster R-CNN where detection is limited to bounding boxes. The Mask R-CNN was also implemented to accurately recognize the HVAC ducts in construction sites with the training time between 4-5 hours. The model results that use transfer learning and data augmentation techniques have a mAP of 90.6%, a precision of 80.87%, and a recall of 87.84%. According to the obtained results , the Mask R-CNN model can accurately detect HVAC ducts with irregular shapes. The main contribution of this research is development of a solution and scheme that can automate HVAC ducts recognition, and its later use in automated progress reporting making use of as-planned and as-built stages via digital imaging taken from either 3D models or real construction jobsites. This will considerably reduce the manual effort and time in monitoring and reporting. To evaluate the performance of our proposed methodology in the construction phase with real HVAC ducts, it is planned to extend our work by utilizing real images. Also in future work, performance of other CNN architectures with more building classes will be fully explored.

7 References

[1] Hamledari H., McCabe B., and Davari S. Automated computer vision-based detection of components of under-construction indoor partitions. *Automation in Construction*, 74:78–94, 2017.

[2] Kopsida M., Brilakis I. and Vela P. A Review of Automated Construction Progress and Inspection Methods. In *Proceedings of the 32nd CIB W78 Conference 2015*, pages 421–431, Eindhoven, The Netherlands, 2015.

[3] Moselhi O., Bardareh H., and Zhu Z. Automated data acquisition in construction with remote sensing technologies. *Applied Sciences*, 10(8):2846, 2020.

[4] Hou X., Zeng Y., and Xue J. Detecting Structural Components of Building Engineering Based on Deep-Learning Method. *Journal of Construction Engineering and Management*, 146(2): 04019097, 2020.

[5] O’Mahony N., Campbell S., Carvalho A., Harapanahalli S., Hernandez GV., Krpalkova L.,

Riordan D. and Walsh J. Deep Learning vs. Traditional Computer Vision. In *Proceedings of the Science and Information Conference*, pages 128–144, Las Vegas, USA, 2019.

[6] Xu S., Wang J., Wang X. and Shou W. Computer vision techniques in construction, operation and maintenance phases of civil assets: A critical review. In *Proceedings of the 36th International Symposium on Automation and Robotics in Construction (ISARC)*, pages 672-679, Banff, Canada, 2019.

[7] Wang Z., Zhang Q., Yang B., Wu T., Lei, K., Zhang B. and Fang T. Vision-Based Framework for Automatic Progress Monitoring of Precast Walls by Using Surveillance Videos during the Construction Phase. Journal of Computing in Civil Engineering, 35(1): 04020056, 2021.

[8] Kim J., Hwang J., Chi S. and Seo J. Towards database-free vision-based monitoring on construction sites: A deep active learning approach. Automation in Construction, 120:103376, 2020.

[9] Ying HQ. and Lee S. A mask R-CNN based approach to automatically construct As-is IFC BIM objects from digital images. In *Proceedings of the 36th International Symposium on Automation and Robotics in Construction (ISARC)*, pages 764-771, Banff, Canada, 2019.

[10] Pour Rahimian F., Seyedzadeh S., Oliver S., Rodriguez S. and Dawood N. On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning. *Automation in Construction*, 110: 103012, 2020.

[11] Kolar Z., Chen H. and Luo X. Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. *Automation in Construction*, 89:58-70, 2018.

[12] Roh S., Aziz Z., and Peña-Mora F. An object-based 3D walk-through model for interior construction progress monitoring. *Automation in Construction*, 20(1):66-75, 2011.

[13] Raoofi H. and Motamedi A. Mask R-CNN Deep Learning-based Approach to Detect Construction Machinery on Jobsites. In *Proceedings of the 37th International Symposium on Automation and Robotics in Construction (ISARC)*, pages 1122-1127, Kitakyushu, Japan, 2020.

[14] Fang W., Ding L., Luo H., and Love PE. Falls from heights: A computer vision-based approach for safety harness detection. *Automation in Construction*, 91:53-61, 2018

[15] Lee A. Comparing Deep Neural Networks and Traditional Vision Algorithms in Mobile Robotics. Swarthmore College, 2015

[16] Wang J., Ma Y., Zhang L., Gao RX. and Wu D. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48:144-156, 2018.

[17] Zou J. and Kim H. Using Hue, Saturation, and Value Color Space for Hydraulic Excavator Idle Time Analysis. *Journal of computing in civil engineering*, 21(4): 238-246, 2007.

[18] Hui L., Park MV. and Brilakis I. Automated Brick Counting for Façade Construction Progress Estimation. *Journal of Computing in Civil Engineering*, 29(6): 04014091, 2015.

[19] Memarzadeh M., Golparvar-Fard M. and Niebles JC. Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, 32: 24-37, 2013.

[20] Nath ND. and Behzadan AH. Deep Learning Models for Content-Based Retrieval of Construction Visual Data. In *Proceedings of the Computing in Civil Engineering*, pages 66–73, Atlanta, USA, 2019.

[21] Tremblay J., Prakash A., Acuna D., Brophy M., Jampani V., Anil C., To T., Cameracci E., Boochoon and Birchfield S. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 969-977, Salt Lake City, USA, 2018.

[22] Ros G., Sellart L., Materzynska J., Vazquez D. and Lopez AM. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, Las Vegas, USA, 2016.

[23] Golkhoo F. *Material Management Framework based on Near Real-Time Monitoring of Construction Operations*. Concordia University, Montreal, 2020.

[24] National Institute of Building Sciences (NIBS). BuildingSMART alliance - common building information model files and tools.On-line: https://www.nibs.org/?page=bsa_commonbimfiles. Accessed: 31/07/2020

[25] Girshick R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, Santiago, Chile, 2015.

[26] Kim C., Son H., and Kym C. The effective acquisition and processing of 3D photogrammetric data from digital photogrammetry for construction progress measurement. *Computing in civil engineering*, 178–185, 2011.

[27] Abdulla W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow.

GitHub repository. On-line: https://github.com/matterport/Mask_RCNN, Accessed: 2/2/2021.

[28] Padilla R., Netto SL. and Da Silva EA. A Survey on Performance Metrics for Object-Detection Algorithms. In *Proceedings of the Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, Niteroi, Brazil, 2020.

[29] Everingham m., Van Gool L., Williams CK., Winn J. and Zisserman A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit.On-line: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/htmldoc/index.html., Accessed: 06/05/2021.