# Image segmentation of underfloor scenes using a mask-RCNN with two-stage transfer learning

Gary A. Atkinson, Wenhao Zhang, Mark F. Hansen

*Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England*
*Bristol, BS16 1QY UK.*

Mathew L. Holloway, Ashley A. Napier

*Q-Bot Ltd., Block G, Riverside Business Centre, Wandsworth, SW18 4UQ UK*

**Abstract**

Enclosed spaces are common in built structures but pose a challenge to many forms of manual or robotic surveying and maintenance tasks. Part of this challenge is to train robot systems to understand their environment without human intervention. This paper presents a method to automatically classify features within a closed void using deep learning. Specifically, the paper considers a robot placed under floorboards for the purpose of autonomously surveying the underfloor void. The robot uses images captured using an RGB camera to identify regions such as floorboards, joists, air vents and pipework. The paper first presents a standard mask regions convolutional neural network approach, which gives modest performance. The method is then enhanced using a two-stage transfer learning approach with an existing dataset for interior scenes. The conclusion from this work is that, even with limited training data, it is possible to automatically detect many common features of such areas.

*Keywords:* Computer vision, underfloor maintenance, convolutional neural network

## 1. Introduction

Many tasks in the built environment require surveys of tightly enclosed or otherwise difficult to access and/or navigate spaces and where data capture is difficult and sparse. Examples range from common domestic properties that require maintenance in roof voids or under floorboards, through to hazardous environments such as nuclear reactors or remote pipework. Such applications require robotic surveyance tools that have minimal footprint in order to access such environments and navigate therein. Any sensors therefore must have minimal dimensions and preferably use low-to-moderate power consumption, such as a standard RGB camera, as in this paper. A further difficulty relates to the time consuming nature of manually controlling such robots to recover required survey data given limited views, awkwardness of terrain and/or the need for laborious data annotation (as is required with deep learning methods and to meet commercial requirements such as accreditation).

This paper offers a method to ease the burden faced in these cases by means of a combined robotics and computer vision system that can rapidly and (semi-) autonomously capture data and provide detailed annotation of the area data. Basic mapping and path-planning is a well-studied area of robotics, which typically involves some form of simultaneous localisation and mapping (SLAM) [1]. However, automated methods to annotate such map data is less documented for the challenging environments considered here. The proposed approach uses deep learning to automatically annotate RGB images from a robot-mounted camera for one particular application: that of underfloor mapping for residential properties. The application is a case study and could be extended to other application areas given sufficient training data.

The application considered here is motivated by a need of Q-Bot Ltd [2] in their underfloor insulation service. A method for automatically spraying insulation into the void is sought where the automatic pixel-wise object classification will enable insulation to be applied where needed to the floorboards and hot water pipes, but prevent insulating over air vents, electrical wiring and joist ends. This task presents a significant challenge even for an experienced human operator, while accreditation requirements create an additional burden (for instance, the British Board of Agrément means the whole installation process must be recorded and documented). Therefore, being able to locate and tag features automatically would streamline an otherwise laborious and repetitive task. These challenges are common across the construction industry. Future applications, using a different set of classes, might involve using robots to apply materials to other building surfaces, using a drone or crawling robot to check facades on a tower block, the detection of cracks in pipework using robots navigating the inside the pipe, or seeking damage to control rods in a nuclear reactor. A new set of labelled training data will be required for these applications but the basic paradigm will be similar to that discussed here.

The specific contributions of the paper are as follows:

- Performance analysis of a general mask *regions with convolutional neural network* (RCNN) approach for semantic segmentation of relevant scenes showing modest performance.
- Methods to improve the robustness of the method using a two-stage transfer learning technique based on an existing dataset and redefinition of feature aspect ratios within the RCNN.
- Detailed experiments to prove the robustness of the method showing strong performance with features such as walls, joists and floorboards and promising results for most others.

Section 2 of this paper considers other literature in the field including the background material from machine learning and neural network methods. Section 3 describes the basic network architecture chosen for this work and the data capture process. Optimisation of the approach by means of two-stage learning and the incorporation of an alternative dataset is covered in Section 4. A detailed analysis of the methods is then provided in Section 5 before Section 6 concludes the paper.

## 2. Related work

The construction industry primarily relies on manual labour-intensive processes and has experienced little productivity growth over the last 25 years, being slow to adopt new technologies [3]. In many countries, the situation is further exacerbated by a skills shortage, which may worsen in the near future [4]. Further, a huge proportion of buildings that will exist in the mid/late 21st century have already been constructed, and to varying environmental standards. For example, it is estimated that 80% of UK buildings that will exist in 2050 have already been built [5]. Therefore, maintaining and upgrading the existing building stock is a key challenge, yet there is very little useful information available to inform decision making and maintenance is often reactive, laborious and difficult.

One reason for the lack of useful information is that many processes rely on paper-based manual data entry (an issue partially addressed in this paper via automated data labelling), and these systems are poorly integrated with business processes. This has created a wide range of challenges as progressively more data is collected, providing a range of new applications for the field of computer vision in the construction industry. Aside from the application of this paper, examples include: partial automation of a tunnel inspection routine [6]; health and safety monitoring [7]; productivity analysis [8]; workforce analysis [9]; detection of interior partition components [10]; recognition and quantification of bugholes in concrete [11]; and extrusion quality monitoring [12].

Despite the increased interest mentioned above, there has been relatively little work to apply computer vision to the many tight spaces found in built structures such as underfloor voids, pipes and around eaves. Some of the more established works in the area involve more predictable conditions with specific requirements, such as pipework [13], and may require some form of structured lighting and/or expensive hardware [14]. In this paper by contrast, we focus on methods for a more generic understanding of tight spaces, with underfloor voids as the case study. Iwaki et al. [15] considered a basic SLAM-based approach to underfloor mapping while a robust version based on similar underlying principles was proposed by Cebollada et al. [16] to apply underfloor insulation [17]. However, these efforts focused on 3D mapping without concern for semantic scene understanding/object detection, which was carried out for this paper.

In line with most automated scene understanding processes in recent years, this work draws on convolutional neural networks (CNNs) [18] due to their proven track record in a range of applications. A neural network is an approach to pass images (or other signals) through a complicated series of filters (convolutions) with particular parameters, such that certain characteristics of the images are enhanced or suppressed. If the correct filters and parameters are chosen, then distinctive features will be enhanced enabling recognition/classification/detection of a particular class of object/scene. Many existing algorithms are able to optimise the parameters for the filters automatically, which yield exceptionally high classification accuracy for numerous applications. However, these methods generally require huge datasets of typically manually labelled images to train for a new application and may require immense computer processing time on graphical processing units (GPUs).

As with many other works in the field, the requirement for huge datasets is partially overcome here using the concept of *transfer learning* [19]: that is, to use a network originally trained on another dataset and retrain certain *layers* of the network to the new application. Nevertheless, the need for sizable datasets is not entirely diminished.

For many applications in construction and elsewhere (e.g. the automated insulation robot considered in this paper), it is insufficient to merely classify a single image, but rather to identify regions within a scene. Girshick et al. proposed a means known as “regions with CNN” (RCNN) to efficiently represent regions within an image in order to localise objects within view [20] forming a bounding box around required targets. This was expanded upon in Fast RCNN and Faster RCNN [21]. The Mask-RCNN [22] provides a pixel-by-pixel mask within the bounding box to form a so-called semantic segmentation but requires an exceptionally time consuming, and usually manual, pixel-wise labelling process in order to make training possible.

The RCNN approaches for semantic segmentation were used for this paper for feature detection. There have been other works applying a related paradigm to the construction industry such as Dung and Anh who semantically segment images of cracks in concrete [23], and Xiong et al. who segment 3D building models from laser scanner point clouds [24]. However, ours is unique in its application to enclosed spaces.

This paper addresses one of the key challenges facing the construction industry: how to improve quality and accountability while accommodating for the lack of existing information. The research shows how computer vision can be used to automate categorisation while accounting for the limited training data set encountered in this application, and many others involving tight spaces, in the construction industry. The main unique features are the novel dataset and tuned RCNN from the enclosed underfloor regions and the nature of the two-stage approach to optimise performance on a limited dataset, while drawing upon a prior model (in this case from the NYU dataset [25]).

## 3. Segmentation using a basic mask-RCNN

This section describes a direct application of an existing Mask-RCNN deep learning architecture to RGB images of underfloor scenes for semantic segmentation.

### *3.1. Data capture and pre-processing*

For the application considered for this paper, a robot that was designed and developed at Q-Bot Ltd (see Fig. 1) was fitted with a standard RGB camera and 2D LiDAR scanner. The LiDAR scanner is not necessary for the bulk of the work in this paper, but is used later to demonstrate the mapping of data labelling to 3D map data. The robot was placed in several underfloor voids of residential properties and navigated remotely throughout the void while continuously capturing image data from multiple viewpoints at regularly spaced angles relative to the robot's direction of travel. Illumination was limited to the on-board white LED sources (see Fig. 1) and a very small amount of natural light from any openings. Further details of the data capture process can be found in [26]. RGB images taken from the camera were then manually labelled and

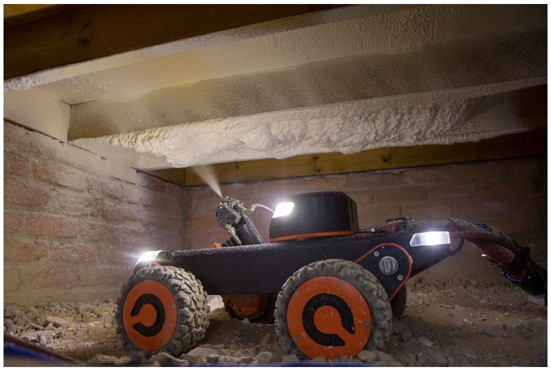


Figure 1: Photograph of the robot used for data capture, shown here applying insulation foam to illustrate an application. The robot dimensions are approximately 58cm(L), 45cm(W), 29cm(H).

used to train an RCNN for segmenting any future images into classes of objects from the environment (e.g. floorboards, walls, joists, etc.). In other words, the goal was to develop a neural network able to accept an RGB image as input and then to output a matrix, of dimensions equal to the original image, where each element indicates the class of object in a given pixel. Fig. 2 illustrates an example of a (near) perfect output matrix produced manually.

A decision to *only* use two-dimensional RGB data (i.e without incorporation of depth) was made since: (1) image-based algorithms are simpler and more efficient; (2) there is no registration necessary between the data from different sensors; (3) the integration of the computer vision code with the rest of a robot coding is more modular; and (4) training data based on transfer-learning (see Section 2) are more readily available in RGB. Nevertheless, it is acknowledged that 3D data does have the potential to offer richer information and will be considered for future work.

The method proposed here takes a raw image captured by the camera and an XML data file containing the manual labels which were established using *LabelMe* [27]. It then creates a mask image containing labels and regions of each object for all the images, similar to that shown in Fig. 2. The data is

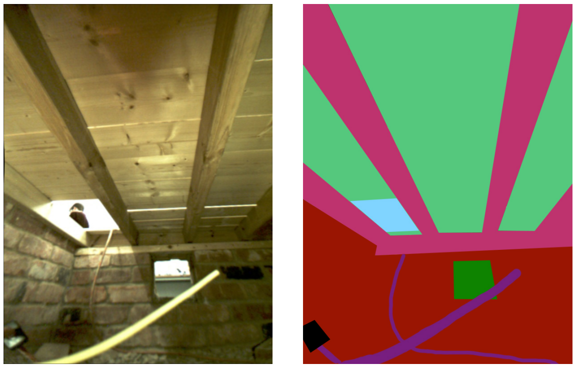


Figure 2: Example of a typical underfloor scene and the theoretical perfect output from the algorithm. The colours encode floorboards (light green), joists (magenta), walls (red), vents (dark green), cables (purple) and openings (cyan). Black areas are unclassified.

split 2:1 between training and testing datasets, as shown in Table 1. The Mask-RCNN is then trained on the training dataset using transfer learning and then validated against the (unseen) testing dataset. The transfer learning is based on pre-trained weights from the *common objects in context* (COCO) dataset [28], since this remains the largest and best-known pixel-wise labelled dataset. The overall framework is illustrated by Fig. 3.

### *3.2. Training data*

A total of 256 labelled images were used for training and evaluating the Mask-RCNN. The breakdown of images and pixels containing the various classes can be seen in Fig. 4. Floorboards, joists and walls appear in the most images and occupy the most pixels.

When the network is trained, half of the data (selected at random) is augmented to simulate a larger dataset by applying the following transformations to generate new data:

- ±10° rotation
- 0.2 of the width/height translation (black padding)
- 50% flipped horizontally

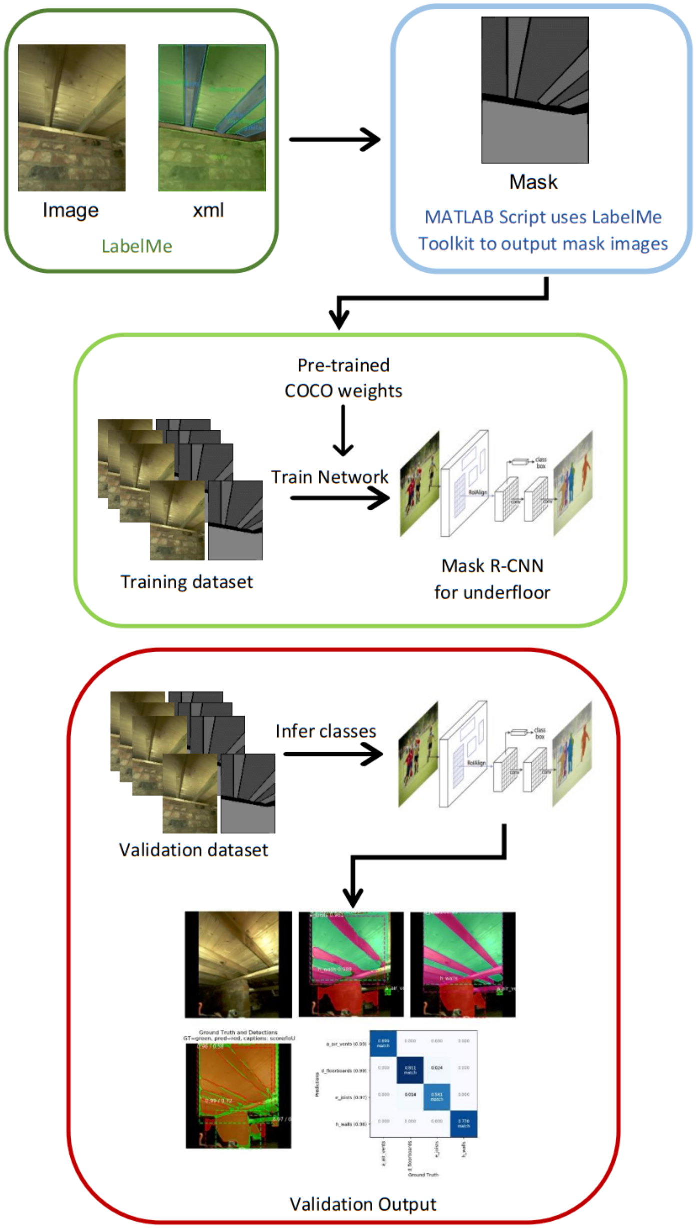


Figure 3: The Mask-RCNN processing pipeline showing the main steps of the process. N.B. the example validation output is for illustrative purposes only, and larger versions can be seen later in the paper (e.g. Fig 5).

These parameters where chosen in light of typical operating conditions of the robot (e.g. where driving over bricks or debris might form an effective small rotation) or geometrical properties of the features (e.g. any floorboard image is equally valid if flipped horizontally but not vertically).

| | Images |
|---|---|
| Training | 170 |
| Testing | 86 |
| Total | 256 |

Table 1: The partitioning of data between training and test datasets.

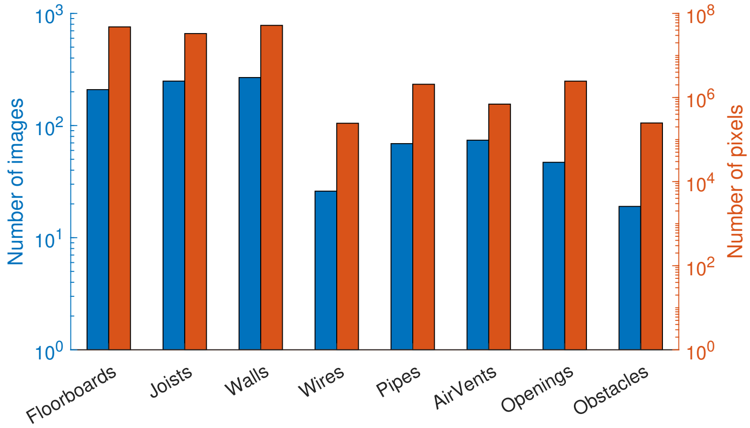


Figure 4: Histogram showing the number of images and number of pixels containing each label. Note the log scale: i.e. there are substantially fewer cases of air vents, obstacles, etc.

### *3.3. Results*

The results described in this section refer to the prediction accuracy of the trained model against the testing partition of the dataset. These are images that have not been used to train the model, but have an associated ground truth mask which is used to assess the performance. The model is applied to each of the 86 images in the testing set, which output an evaluation graphic for each test as shown in Fig. 5. These show:

- The input image.
- The predictions of the model overlaid on the image.
- The ground truth labels.
- An overlay image of the predictions and ground truth, where red shows the predicted mask and green shows the ground truth.
- The confusion matrix of ground truth vs predicted labels (e.g. 66.2% of floorboard pixels are correctly identified as floorboards for the case of Fig. 5).

Figure 6 shows the accumulated results for all classes across all of the 86 images in the testing set (i.e. it is the output if all the confusion matrices generated for each image are combined). This demonstrates that fair recognition accuracy for walls (86.3%) and floorboards (73.5%) has been achieved. The accuracy for a few other categories appears low, which is probably due to the fact that the relative pixel counts become obscure when aggregating all the confusion matrices: a single misdiagnosed large object can have an exaggerated effect on the overall accuracy that is reported here. The low accuracy for some classes is also likely due, in part, to human errors in the data labelling process, as well as the biased distribution of pixel count per category. We note that, if supplied with sufficient accurate training data, as with the case of walls and floorboards, the same model should be able to yield a much higher overall accuracy. Further, individual results such as those in Fig. 5 have the potential for much higher accuracy than the accumulated totals imply. Nevertheless, the results from this basic approach are only modest and so motivated the improvements described in Section 4.

### *3.4. Limitations*

Based on results for this section, the main limitation would appear to be those object types that are small and/or have a low number of examples/pixels. Importantly, these often correspond to wires and pipes which have a small number of occurrences, and also have a high degree of variance associated with their

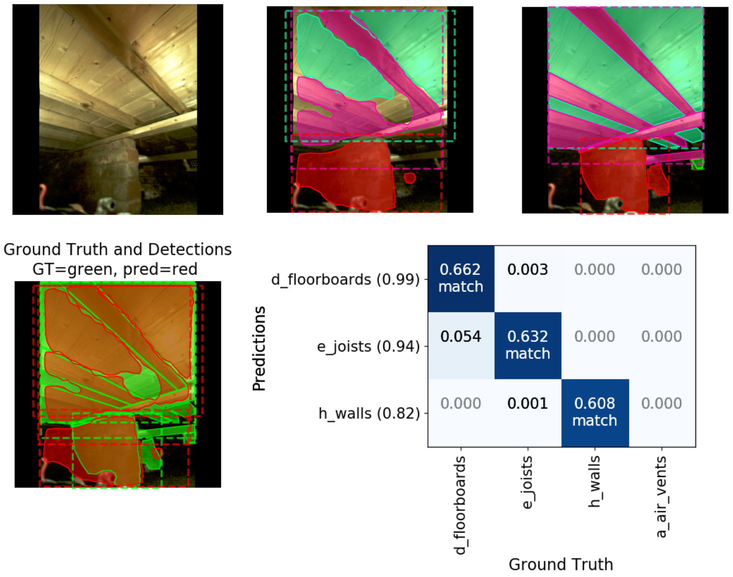


Figure 5: Example results for a typical test image. Top row: input image, prediction labels, ground truth labels. Bottom row: overlaid predictions, confusion matrix of predicted vs actual class labels, where the numbers in brackets following the predicted class labels are probability scores of each label prediction being correct.

appearance. Although these objects are detected in some images, they are so badly identified across the whole dataset, that they all score close to zero.

To help understand this problem, a sequence of real images were augmented with synthetic wires (based on Bezier curves) before the CNN was retrained. With a sufficiently large number of new images, many of the synthesised wires were detected, proving the overall validity of the network architecture for such awkwardly shaped features and motivating the following section. Note that, while the validity of the architecture was proven this way, we later found no evidence that synthetic images helped to detect real pipes or wires and so were not included in further training sets.

Aside from the issue of small/less common features, a major weakness of the method above relates to the boundaries of regions. This is apparent in Fig. 5, where a portion of floorboards near the centre of the image has not been segmented into anything at all.

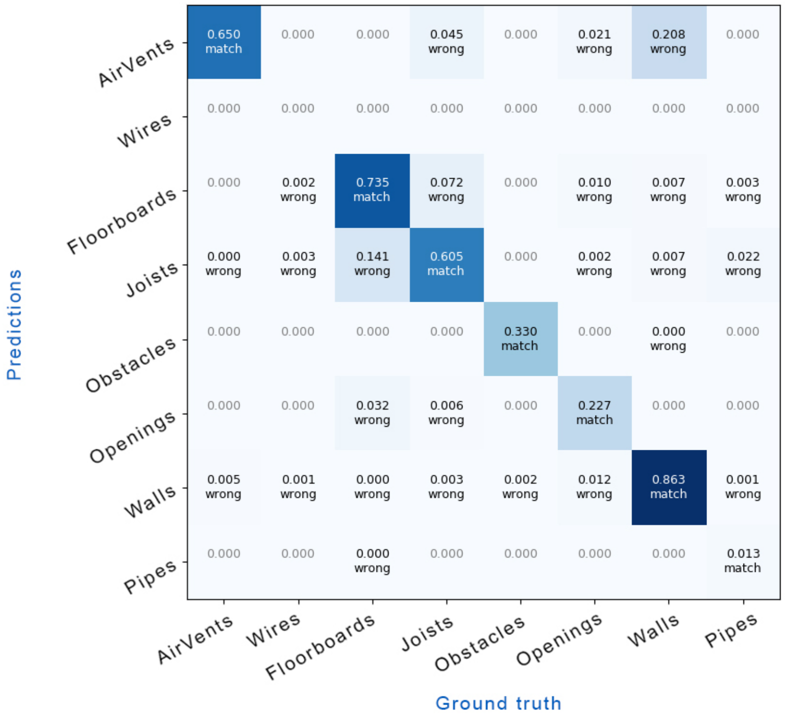


Figure 6: Performance matrix using all 86 images in the testing partition.

## 4. Segmentation using a fine-tuned two-stage mask-RCNN

This section covers two methods by which the training of the network was optimised: using a two-stage learning approach and considerations of the aspect ratio (anchor) of detection regions.

### *4.1. An improved method*

One of the greatest challenges faced in this project was the difficulty in capturing enough good-quality image data to train a reliable Mask-RCNN model; especially for certain classes, such as pipes and wires. Therefore, a two-stage transfer learning approach is proposed to learn a "transitional model" for the semantic segmentation of objects in buildings, and then to the final model for semantic segmentation in underfloor voids. In the two-stage learning process, the model benefits from being exposed to more pipe and wire data and it is therefore capable of retaining useful features learned from these object classes. It is then tuned to work with data captured in underfloor voids and use these features for an improved classification and segmentation performance. This method ensures that the majority of knowledge (i.e. network weights) that is transferred from the first stage to the second stage is useful and relevant.

To obtain the transitional model, a subset of the New York University (NYU) dataset [25] was employed which contains image data of walls, floors, ceilings, pipes, wires, air vents, etc., similar to the objects of interest in this project. The dataset also contains unrelated classes such as paintings, desks and cabinets. Overall, 1449 images containing over 10,000 instances were utilised for this first-stage learning. *All* model layers were trained with a batch size of 1 for 40 epochs, at which time it converged to its final values. 80% of the NYU subset was used for training and the remaining 20% for testing. For this first stage, many general classes were used, including those not directly related to underfloor scenes. This approach helps to capture low-level features (such as corners and edges of general indoor objects) from a larger dataset than would be possible using underfloor features alone. These features are potentially useful for the second stage of learning, even though they originate from unrelated classes.

The trained model weights were then used for a second stage transfer learning, during which only the model heads were trained: i.e. only the output layer, not the region proposal network nor the backbone model. Given that there were fewer trainable model parameters than those in the previous stage, a batch size of 2 was used. The model was trained for 85 epochs before it started to over-fit. Note that the same pre-trained COCO weights were used for model initialisation as those in the method of Section 3. Further, the same 86 images in the testing partition were used to evaluate this improved model such that performance of this model can be directly and objectively compared to the original model. Other details of the training process are the same as illustrated in Section 3 while the overall framework is illustrated by Fig. 7.

Another investigation was carried out with a focus on pipe and wire detection/classification. Pipes and wires are often found to be elongated in the

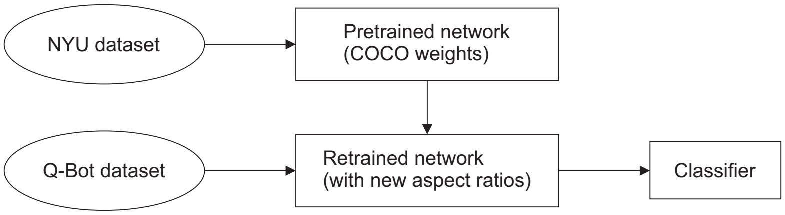


Figure 7: Framework for the two-stage transfer learning approach.

captured images. However, the original Mask-RCNN model generates Region-of-Interest (ROI) proposals (i.e. bounding boxes for detected objects) that have aspect ratios restricted to 0.5, 1 or 2, meaning that it will be effective in detecting square objects but may struggle with thin objects. Therefore, the aspect ratios were replaced with 0.2, 0.5, 1, 2 and 5, and the region proposal network layers of the model retrained to accommodate the new aspect ratios.

### *4.2. Results*

Figure 8 shows results using the new method for the same scene as that shown in Fig. 5. The result clearly shows a cleaner transition between the segments and better coverage of the floorboards.

Figure 9 shows the updated confusion matrix using the improved training method. The matching scores are clearly higher than in Fig. 6 except for the "obstacles" class. The reason for this is uncertain although fortunately this is likely to be a less critical class than most others since obstacles can easily be identified using 3D scanning hardware rather than CNN approaches.

Further investigation showed that improvements between results in Figs. 6 and 9 came from both the modification of aspect ratios and the inclusion of NYU data. This is exemplified by Table 2, which shows the *detection* rate (i.e. fraction of instances in which a class is identified at all, as opposed to the amount of region overlap) for pipes increased substantially as a result of both measures. This is despite the IoU scores being modest for all approaches. This is significant since it is often adequate to detect the presence (or otherwise) of a particular feature without details of the precise pixel distribution.

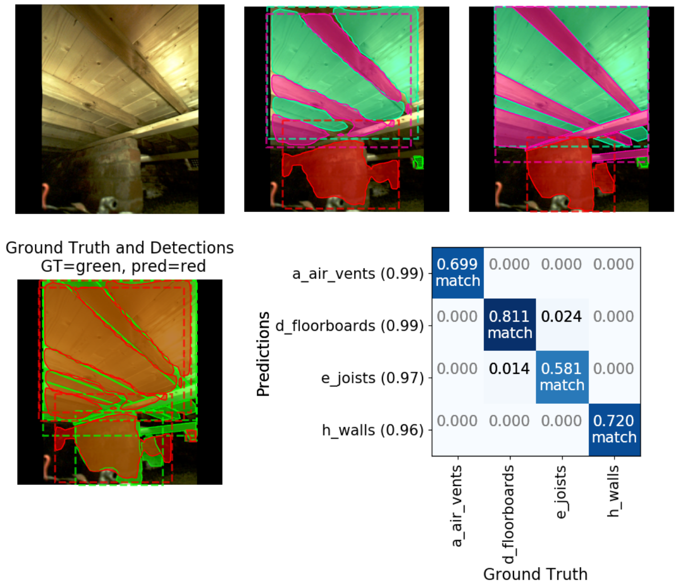


Figure 8: Improved segmentation of the scene from Fig. 5.

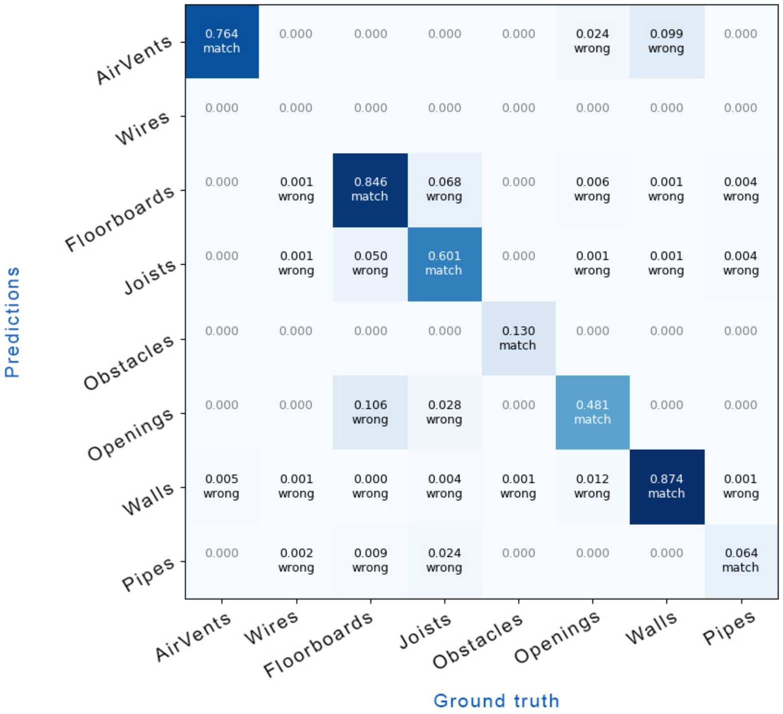
| | (a) | (b) | (c) |
|---|---|---|---|
| Pipes (out of 23) | 4 (17.4%) | 9 (39.1%) | 18 (78.3%) |
| Wires (out of 6) | 0 | 2 (33.3%) | 2 (33.3%) |

Table 2: Detection rates for pipes and wires using (a) the basic method from Section 3, (b) the improved method with different aspect ratios, and (c) the final approach including the two-stage learning.

## 5. Summary and discussion

To understand the improvements/differences between the basic approach and the advanced approach, one should consider the confusion matrices (the diagonals of which are summarised more clearly in Fig. 10), the detection rates from Table 2 and Fig. 11, and the individual scene performances.

The general comparison, illustrated by Fig. 10, shows that results for most classes are either comparable or notably better for the advanced method. While pipe and wire scores are very low indeed for both methods, Table 2 shows better detection rate for the advanced approach, as already mentioned. Figure 11 shows the detection rates for all classes, mostly revealing a similar pattern to



| | AirVents | Wires | Floorboards | Joists | Obstacles | Openings | Walls | Pipes |
|---|---|---|---|---|---|---|---|---|
| AirVents | 0.764<br>match | 0.000 | 0.000 | 0.000 | 0.000 | 0.024<br>wrong | 0.099<br>wrong | 0.000 |
| Wires | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Floorboards | 0.000 | 0.001<br>wrong | 0.846<br>match | 0.068<br>wrong | 0.000 | 0.006<br>wrong | 0.001<br>wrong | 0.004<br>wrong |
| Joists | 0.000 | 0.001<br>wrong | 0.050<br>wrong | 0.601<br>match | 0.000 | 0.001<br>wrong | 0.001<br>wrong | 0.004<br>wrong |
| Obstacles | 0.000 | 0.000 | 0.000 | 0.000 | 0.130<br>match | 0.000 | 0.000 | 0.000 |
| Openings | 0.000 | 0.000 | 0.106<br>wrong | 0.028<br>wrong | 0.000 | 0.481<br>match | 0.000 | 0.000 |
| Walls | 0.005<br>wrong | 0.001<br>wrong | 0.000<br>wrong | 0.004<br>wrong | 0.001<br>wrong | 0.012<br>wrong | 0.874<br>match | 0.001<br>wrong |
| Pipes | 0.000 | 0.002<br>wrong | 0.009<br>wrong | 0.024<br>wrong | 0.000 | 0.000 | 0.000 | 0.064<br>match |

Figure 9: Performance matrix using all 86 images in the testing partition using the improved training regime.

that already found. It should be noted that the detection rates are manually provided on a *per image* basis and involve a certain amount of subjectivity. For example, the bottom-right image in Fig. 12, is classed as *floorboards not detected* since there is a floorboard present that is not found at all.

A qualitative assessment by visual reference to individual scenes reveals little obvious trend in terms of resilience to illumination, perspective effects, etc. Indeed, there are some cases whereby the basic method performs better for unknown reasons. Consider the results in Fig. 12 for example. The first result shows a challenging case where the advanced method performs better, as expected, while the second case involves one joist that was only detected in the basic method.

Figures 13 to 15 show further results for the advanced method to illustrate

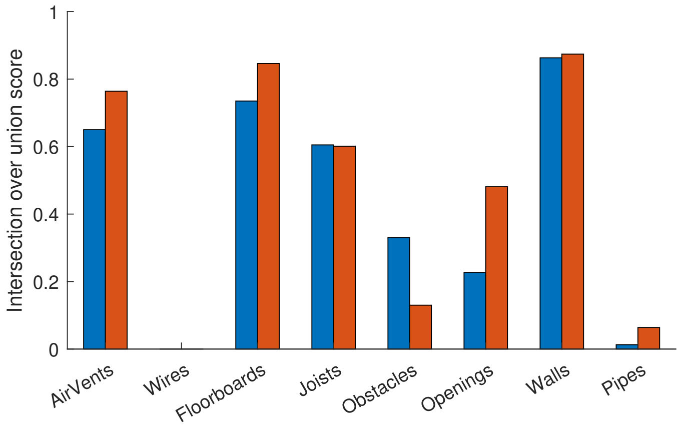


Figure 10: Intersection over union scores for the basic method (left bars for each class) and the advanced method. The values here correspond to the diagonals of Fig. 6 and 9.

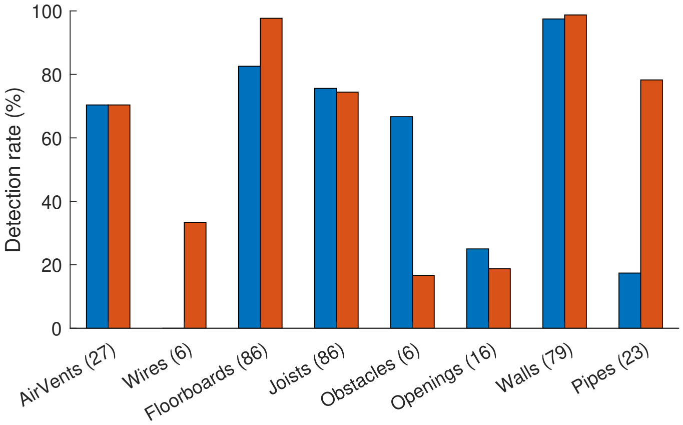


Figure 11: Detection rates for the basic method (left bars for each class) and the advanced method. Numbers in brackets refer to the the number of images containing that class.

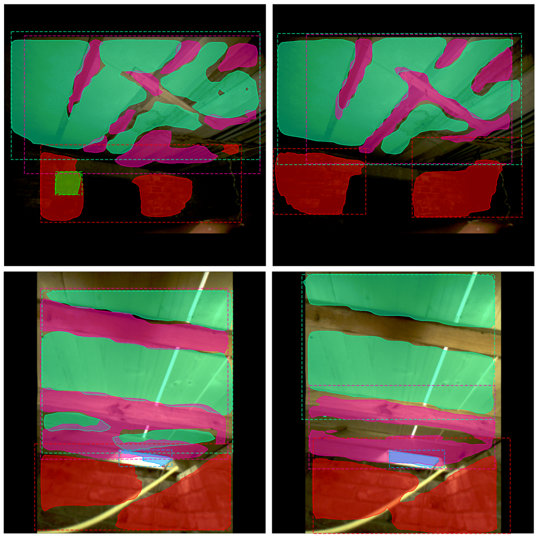


Figure 12: Results from the basic method (left) and advanced method for two typical scenes.

its strengths and limitations. The first two of these show very successful segmentation. The last example is more challenging due to the complex patterns of walls, joists and pipes. However, it is only the pipework that was completely undetected in this case.

Finally, Fig. 16 shows an example of mapping the classified areas to 3D data for applications in robotics: e.g. to determine where to apply insulation and what application stroke technique to apply. In real applications, it may be necessary to further supplement this with classification scores from the detection stage of the network. This would help to avoid a misclassification: for example, if the vent to the lower-right of Fig. 16 were detected with low confidence, then a human operator could be called upon for confirmation.

## 6. Conclusions

This paper has presented a CNN approach to the highly challenging task of automated underfloor closed-space scene segmentation. While the various performance metrics may not suffice for all applications, there is strong evidence that the CNN can reliably segment scenes given sufficient training data.

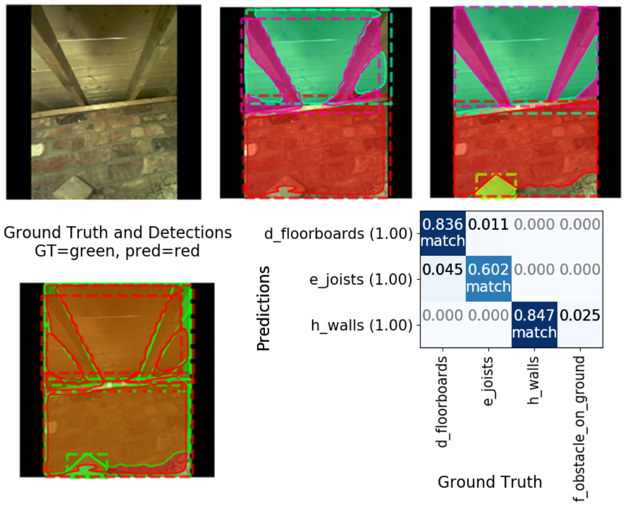


Figure 13: Example of a well-segmented simple scene.

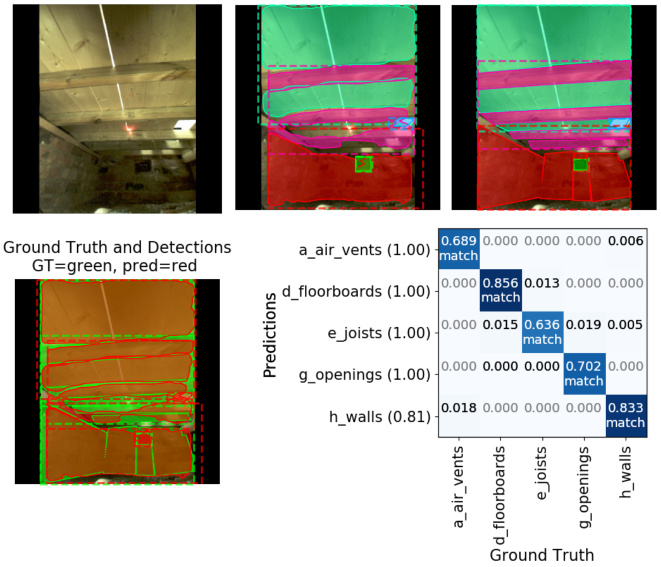


Figure 14: Example of a more challenging well-segmented simple scene.

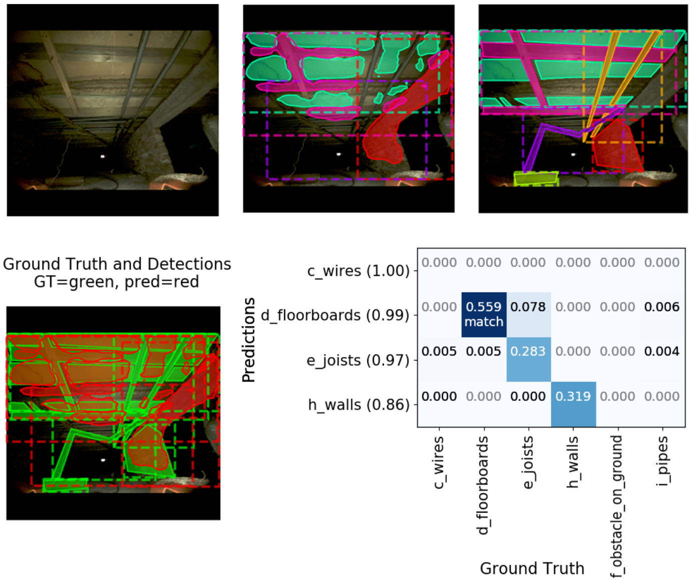


Figure 15: Example of challenging scene.

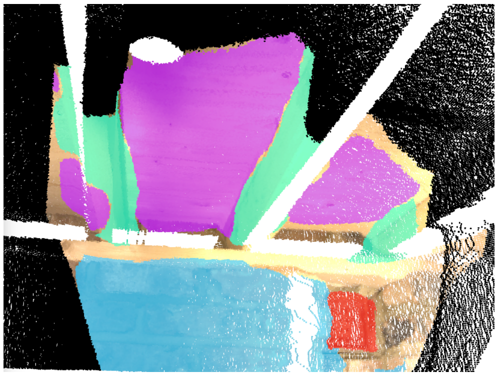


Figure 16: Example of mapping the 2D labels onto a region of a 3D point cloud (here LiDAR data).

The work constitutes a step forward towards remote and autonomous labelling of features in closed environments. This is because the developed methods are robust for many of the most ubiquitous and important features; the methods can be deployed in a real system; and the proposed paradigm offers a framework that can be augmented in future application areas. The fine-tuning of the algorithm for objects with high aspect ratios is of particular significance to the application of computer vision in construction: many common construction objects are long and thin, including those not specifically studied in this paper such as truss structures, scaffold and cranes. While these features were not part of this study, it is expected that a very similar approach could be applied for their detection. The paper also demonstrates that the NYU dataset is appropriate for follow-on research where segmentation of construction environments/built scenes is essential (either using the approach in this paper or otherwise). Finally, it should be reiterated that all deep learning methods require large datasets. Therefore, appending any future manually labelled images to the dataset and retraining the network should improve performance further – as would the incorporation of any depth sensor data.

This research has demonstrated the potential for computer vision techniques to impact the construction industry in two areas:

1. Improving the automation of robotic systems: in this case allowing a robot to identify areas needing treatment and allowing for the generation of an appropriate control strategy.
2. Creating "digital twins", or annotated records, of buildings where none currently exist.

This creates opportunities to improve planning and management of a wide range of construction activities, enhances the information available for quality assurance, as well as improving the productivity of the workforce. For Q-Bot's application specifically, the information is critical for accreditation; creating an efficient, consistent (as the process is not reliant on an individual's interpretation and record keeping) and accountable process to verify installation quality.

**Acknowledgements**

This work was funded by InnovateUK (TSB Reference: TS/P010954/1).

## References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, J. J. Leonard, Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age, IEEE Transaction on Robotics 32 (6) (2016) 1309–1332. doi:https://doi.org/10.1109/TRO.2016.2624754.

[2] Q-Bot Ltd., https://www.q-bot.co/, [Accessed: 10 January 2020].

[3] UK government white paper, "Fixing our broken housing market", https://www.gov.uk/government/publications/fixing-our-broken-housing-market, [Accessed: 10 January 2020].

[4] Homes England Strategic Plan, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/752686/Homes_England_Strategic_Plan_AW_REV_150dpi_REV.pdf, [Accessed: 10 January 2020].

[5] The Royal Society of Engineering, Engineering a low carbon built environment, ISBN 1-903496-51-9, [Accessed: 10 January 2020] (2010).
URL https://www.raeng.org.uk/publications/reports/engineering-a-low-carbon-built-environment

[6] S.-N. Yu, J.-H. Jang, C.-S. Han, Auto inspection system using a mobile robot for detecting concrete cracks in a tunnel, Automation in Construction 16 (3) (2007) 255 – 261. doi:https://doi.org/10.1016/j.autcon.2006.05.003.

[7] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Advanced Engineering Informatics 29 (2) (2015) 239 – 251. doi:https://doi.org/10.1016/j.aei.2015.02.001.

[8] J. Gong, C. H. Caldas, Computer vision-based video interpretation model for automated productivity analysis of construction operations, Journal of Computing in Civil Engineering 24 (3) (2010) 252–263. doi:https://doi.org/10.1061/(ASCE)CP.1943-5487.0000027.

[9] E. Konstantinou, J. Lasenby, I. Brilakis, Adaptive computer vision-based 2D tracking of workers in complex environments, Automation in Construction 103 (2019) 168 – 184. doi:https://doi.org/10.1016/j.autcon.2019.01.018.

[10] H. Hamledari, B. McCabe, S. Davari, Automated computer vision-based detection of components of under-construction indoor partitions, Automation in Construction 74 (2017) 78 – 94. doi:https://doi.org/10.1016/j.autcon.2016.11.009.

[11] F. Wei, G. Yao, Y. Yang, Y. Sun, Instance-level recognition and quantification for concrete surface bughole based on deep learning, Automation in Construction 107 (2019) 102920. doi:https://doi.org/10.1016/j.autcon.2019.102920.

[12] A. Kazemian, X. Yuan, O. Davtalab, B. Khoshnevis, Computer vision for real-time extrusion quality monitoring and control in robotic construction, Automation in Construction 101 (2019) 92 – 98. doi:https://doi.org/10.1016/j.autcon.2019.01.022.

[13] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, P. Fieguth, A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure, Advanced Engineering Informatics 29 (2) (2015) 196 – 210. doi:https://doi.org/10.1016/j.aei.2015.01.008.

[14] Laser Tunnel Scanning System, http://www.pavemetrics.com/applications/tunnel-inspection/laser-tunnel-scanning-system/, [Accessed: 10 January 2020].

[15] H. Mizumoto, N. Sata, H. Iwaki, S. Oomura, S. Tsukui, F. Matsuno, Development of intuitive operation interface of underfloor inspection robot, in: Society of Instrument and Control Engineers Annual Conference, 2008, pp. 2962–2967. doi:https://doi.org/10.1109/SICE.2008.4655170.

[16] S. Cebollada, L. Payá, M. Juliá, M. Holloway, O. Reinoso, Mapping and localization module in a mobile robot for insulating building crawl spaces, Automation in Construction 87 (2018) 248 – 262. doi:https://doi.org/10.1016/j.autcon.2017.11.007.

[17] M. Holloway, M. Juliá, P. Childs, A robot for spray applied insulation in underfloor voids, in: In Proc. 47th International Symposium on Robotics, Munich, Germany, 2016, pp. 313–319.

[18] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, Massachusetts Institute of Technology, ISBN 978-0262035613, 2016.

[19] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big Data 3 (2016) 9. doi:https://doi.org/10.1186/s40537-016-0043-6.

[20] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587. doi:https://doi.org/10.1109/CVPR.2014.81.

[21] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2017) 1137–1149. doi:https://doi.org/10.1109/TPAMI.2016.2577031.

[22] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2) (2020) 386–397. doi:https://doi.org/10.1109/TPAMI.2018.2844175.

[23] C. V. Dung, L. D. Anh, Autonomous concrete crack detection using deep fully convolutional neural network, Automation in Construction 99 (2019) 52 – 58. doi:https://doi.org/10.1016/j.autcon.2018.11.028.

[24] X. Xiong, A. Adan, B. Akinci, D. Huber, Automatic creation of semantically rich 3D building models from laser scanner data, Automation in Construction 31 (2013) 325 – 337. doi:https://doi.org/10.1016/j.autcon.2012.10.006.

[25] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: Proc. European Conference on Computer Vision, Florence, Italy, Lecture Notes in Computer Science, Vol. 7576, 2012, pp. 746–760. doi:https://doi.org/10.1007/978-3-642-33715-4_54.

[26] M. Julia, M. Holloway, O. Reinoso, P. R. N. Childs, Autonomous surveying of underfloor voids, in: Proc. 47th International Symposium on Robotics, Munich, Germany, 2016.

[27] Label Me, MIT Computer Science and Artificial Intelligence Lab, http://labelme.csail.mit.edu/Release3.0/, [Accessed: 10 January 2020].

[28] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: Proc. European Conference on Computer Vision, Zurich, Switzerland, Lecture Notes in Computer Science, Vol. 8693, 2014, pp. 740–755. doi:https://doi.org/10.1007/978-3-319-10602-1_48.