# Vision-based Segmentation, Measurement, and Pose Estimation in Modular Integrated Construction

**Yuanyang Qi[1], Xiao Li[2], and Ruiqi Jiang[3]**

[1]Department of Civil Engineering, The University of Hong Kong, China
[2]Department of Civil Engineering, The University of Hong Kong, China
[3]Department of Civil Engineering, The University of Hong Kong, China

yuanyang.qi@connect.hku.hk, shell.x.li@hku.hk, richj233@connect.hku.hk

**Abstract -**

**Modular Integrated Construction (MIC) currently depends on manual assembly processes, which are not only inefficient but also have safety risks. To address this issue, we propose vision-based technologies for automatic segmentation, measurement, and position estimation to aid in the assembly process in construction. Specifically, we utilize the Yolov8-seg model, which is an advanced instance segmentation tool, to segment the surfaces of the corners of the modules. These segmented surfaces are then fitted with quadrilaterals to accurately localize the four corner points. The 3D coordinates of these corners are determined by using the corresponding depth map, thus enabling precise measurements. Additionally, using the 2D coordinates of the colour map of the four corners and the actual 3D coordinates created with the center of the quadrilateral as the origin, we employ the Perspective-n-Point (PnP) algorithm for precise position estimation. The experiments show the effectiveness of the proposed methods. By integrating these vision-based techniques with construction robotics, we can significantly enhance the efficiency and safety of MiC assembly, paving the way towards full construction automation.**

**Keywords -**

**Computer Vision; Segmentation; Measurement; Pose Estimation; Modular Integrated Construction**

## 1 Introduction

Modular Integrated Construction (MIC) involves the design of standardized modules, which are pre-fabricated in off-site manufacturing facilities and subsequently assembled on-site [1]. This method simplified the building process, saving project time and costs. However, the module installation process remains arduous and time-consuming. Lifting large modules requires cranes and hoists, and relies heavily on worker experience to achieve precise lifting; if workers are inexperienced, installation errors, delays, and safety accidents may occur, affecting project time and costs [2]. Due to the increasing demand for faster, safer and more efficient installations [3], the integration of new technologies such as computer vision is becoming important to improve accuracy, speed, and safety in module assembly.

Vision-based techniques have great potential to solve problems in the field of construction, offering a range of advantages that improve the overall construction process, including quality management, progress and schedule monitoring, safety enhancement, and cost reduction [4]. These vision-based methods involve advanced image processing, deep learning, 3D reconstruction, and other techniques to extract valuable visual perceptual information. Furthermore, when integrated with construction robots, these techniques—such as segmentation, measurement, and pose estimation—can significantly improve the robots' perception, enabling more efficient and automated construction.

Segmentation is extracting useful regions from an image or video frame for analysis [5]. In recent years, deep learning applications such as convolutional neural networks (CNN) have greatly improved the accuracy and efficiency of image segmentation [6]. Generally, segmentation can be divided into semantic segmentation and instance segmentation, each provides solutions for different kinds of tasks. Semantic segmentation is the categorisation of each pixel into a specific class, without distinguishing between specific objects in each class. One of the pioneering works of semantic segmentation is Fully Convolutional Networks (FCN) [7], which achieves end-to-end image segmentation by grouping each pixel into specific classes using convolutional layers rather than fully connected layers. U-Net [8], on the other hand, uses an encoder-decoder structure to be able to capture both local and global features and skip connections to fuse deep and shallow information. DeepLab [9] uses dilated convolution to increase the sensory field, which facilitates the segmentation of objects at different scales. Instance segmentation not only focuses on which class each pixel belongs to but also focuses on distinguishing different objects in the same class, which is more challenging than semantic segmentation. Mask R-CNN [10] is one of the most widely used instance segmentation models, which improves on Faster R-CNN [11]

by adding pixel-level segmentation to achieve detection and segmentation at the same time. Leanne Attard et al. [12] employed Mask R-CNN to detect and localize cracks on concrete surfaces, thereby overcoming the limitations of manual detection and significantly reducing both time and cost.

Measurement technology plays a crucial role in Modular Integrated Construction (MIC), particularly for the precise measurement of prefabricated modules. Traditional measurement methods, which rely on manual techniques, are time-consuming, labour-intensive, and subjective [13]. In contrast, vision-based methods such as binocular vision, Time of Flight (ToF), and structured light offer more accurate, objective, and reliable measurement solutions. Binocular vision simulates the human eye by using two cameras to take pictures of the same object at different angles, and by calculating the disparity of the same object in the pictures of the two cameras, thus achieving the depth measurement of the object [14]. ToF obtains depth information by measuring the time it takes for the light signal to travel from the emitting source to the target object and back to the sensor [15]. The structured light system achieves 3d reconstruction of the object by projecting a known pattern onto the target object, and the camera achieves this by capturing changes in the pattern [16]. Liu [17] uses binocular vision to measure the size of cracks in walls, preventing and treating wall defects in advance, thus enabling structural and environmental monitoring of buildings. Yu et al. [18] combined structured light with an industrial robot to stitch together 3d data, simplifying the calibration process and improving the measurement accuracy, thus enabling accurate measurement of the 3D shape of large objects.

Pose estimation is widely used in augmented reality, construction robotics, and industrial automation to enable three-dimensional reconstruction, robotic assembly, and automated construction. Pose estimation aims at obtaining the position and attitude of an object in 3D space, which is crucial for the accurate placement and installation of MIC modules. Attitude estimation methods are classified into feature-based methods and deep learning-based methods etc. Feature-based methods extract feature points such as corner points, edges, key points, etc. from images or video frames and use feature matching to estimate position and orientation. For example, the Perspective-n-Point (PnP) algorithm uses a projection model to estimate pose by extracting feature points from 2d images and matching them with actual 3D coordinates [19]. Mi et al. [20] used the PnP algorithm for 3D position measurement of the pit to achieve automated monitoring of pit displacement. Deep learning-based algorithms use CNN to predict the pose directly end to end. PoseNet [21] based on CNN directly estimates the position and orientation of the camera in 3D space from a single image without feature matching.OpenPose can extract the keypoints from a single image and supports multi-person detection. Prabesh Paudel et al. [22] used PoseNet to detect the worker's position and assess the risk based on the worker's posture to ensure safety. Deep learning-based methods require a large amount of training data to learn effective features with high computational complexity, whereas feature-based pose estimation algorithms require less data, have low computational complexity and good real-time performance.

There are also several studies that apply computer vision techniques to MIC to address challenges in the manual assembly process. For instance, Roshan Panahi et al. [23] proposed an automated assembly progress monitoring system for modular construction factories using computer vision-based instance segmentation. Zhenjie Zheng et al. [24] used Mask R-CNN in the construction process, thus enabling accurate localisation and segmentation of MIC modules, which is conducive to automated process monitoring. However, while these studies focus on segmentation and monitoring, there remains a gap in fully automating not just the detection but also the precise measurement and pose estimation of construction modules.

In this study, we aim to develop a computer vision-based framework for automated segmentation, precise measurement, and position estimation of construction components, with a focus on MIC corner surfaces. The specific objectives of this study are: (1) to achieve accurate instance segmentation of MIC module corner surfaces using the Yolov8-seg model, (2) to extract and fit segmented surfaces into quadrilaterals for identifying precise corner points and measuring the MIC corner, and (3) to implement the PnP algorithm for pose estimation by using the relationship between 2D and 3D corner coordinates.

The contributions of this study include: (1) the integration of state-of-the-art computer vision techniques to automate traditionally manual tasks in the construction industry, (2) a novel application of instance segmentation and 3D depth-based analysis to achieve precise segmentation and measurement MIC components, and (3) the pose estimation of construction components through the combined use of depth data and 2D-3D point correspondence.

The following sections of this paper is structured as follows: Section 2 details the proposed methodology, including segmentation, measurement, and the PnP-based pose estimation process. Section 3 presents experimental results and evaluates the performance of the proposed framework. Finally, Section 4 provides a conclusion, summarizing the key findings.

## 2 Methodology

Figure 1 shows a typical MIC module. It is notable that there is a distinctive rectangular fitting at each of the four

corners of the module. This fitting has unique geometric features and functions that not only provide additional support for the module but also facilitate quick and precise assembly. Therefore this fitting is an important feature of the MIC module. By analyzing the corner features, we can achieve the detection and positioning of the corners of the MIC module, which enables the robot to automatically identify and manipulate the MIC module and improves the automation level of the construction.

Figure 1. Typical MIC module.

To further investigate the corner feature of the MIC module, a 1:1 model of the corner shown in Figure 2 was 3D printed to investigate automated vision-based processing methods, including segmentation, measurement, and pose estimation. The model allows us to simulate and evaluate the performance of the vision system accurately under laboratory conditions.

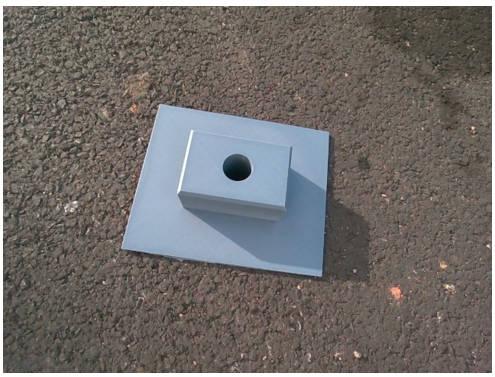

Figure 2. 3D printed 1:1 model of the MIC module corner.

### 2.1 Segmentation

Considering the trade-off between accuracy and processing speed, we have selected the Yolov8-seg model for image segmentation. Yolov8 introduces several enhancements to optimize performance, including an anchor-free detection method, the C2f module, a decoupled header, and a modified loss function [25]. These improvements collectively enhance the model's speed, accuracy, and generalization capabilities. Yolov8-seg, a variant derived from Yolov8, is specifically engineered for segmentation tasks. It inherits the high accuracy and speed of the YOLO series and extends these capabilities to achieve pixel-level segmentation. The architecture of Yolov8-seg is illustrated in Figure 3.

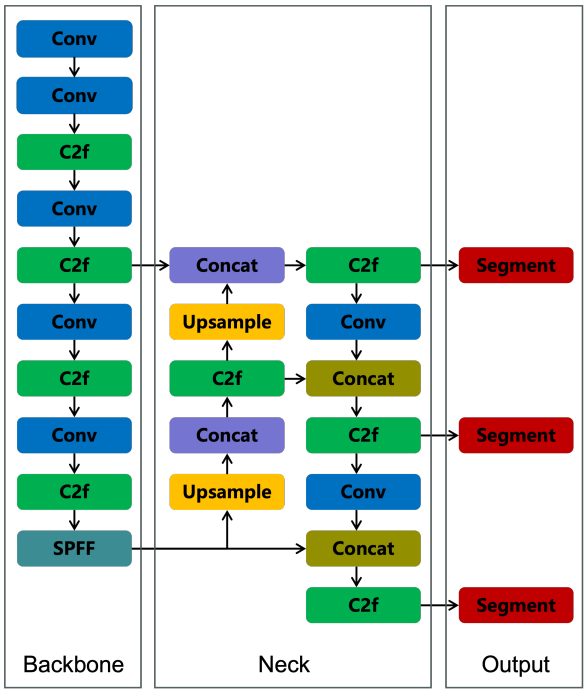


Figure 3. The architecture of Yolov8-seg.

Accurate image segmentation is essential for precise subsequent analysis. In our study, we employed the Yolov8-seg model to segment the surface of the corner model, which allows us to detect and localize the specific positions of the corner model accurately. After segmenting the image, we fit it into a quadrilateral to extract the four corner points. In this way, we can know the exact position of the four corner points in the 2D image and get the 2D coordinates of the four corner points. The segmented image and the coordinates of the four corner points provide a robust foundation for further geometric and dimensional analyses. These will subsequently be utilized for in-depth measurements and accurate pose estimation.

### 2.2 Measurement

The process of mapping points from a three-dimensional coordinate system to a two-dimensional image plane can be described using the pinhole camera model, which is shown in Figure 4.

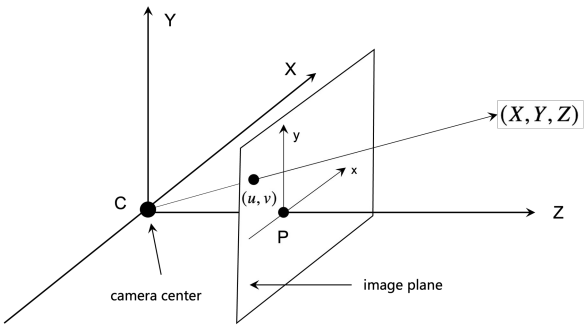


Figure 4. Pinhole camera model.

According to the pinhole camera model, the projection of a point $(X, Y, Z)$ in three-dimensional space onto a point $(u, v)$ on the two-dimensional image plane is described by the following equations:

$$\begin{aligned} u &= f_x \frac{X}{Z} + c_x, \\ v &= f_y \frac{Y}{Z} + c_y, \end{aligned} \tag{1}$$

where $f_x$ and $f_y$ are the focal lengths along the x and y axes, respectively, and $c_x$ and $c_y$ are the coordinates of the principal point on the image plane.

This can be further expressed in matrix form:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \\ 1 \end{bmatrix} \tag{2}$$

where $K$ is the camera's intrinsic matrix and defined as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

To accurately measure the size of the corner model, we use a depth camera. The depth camera captures both a colour map and a corresponding depth map, with the depth map providing the depth value (i.e., the z-coordinate) for the 3D points. Using the formula:

$$\begin{aligned} X &= (u - C_x) \times \frac{Z}{f_x} \\ Y &= \left(v - C_y\right) \times \frac{Z}{f_y} \\ Z &= \text{ depth value} \end{aligned} \tag{4}$$

we can accurately derive the 3D coordinates of any point from its depth map. Using the 2d image coordinates of the four corner points obtained from the segmentation result in the previous step, combined with the depth map, the 3d coordinates of the four corner points can be obtained. The actual length and width of the corner model can be obtained by calculating the distance between two neighbouring points from the 3d coordinates of the four corner points.

### 2.3 Pose Estimation

Since we have previously determined the actual length and width of the corner model, we can construct a 3D coordinate system for the corner model based on the actual length and width. By setting the center point of the model as the coordinate origin, we can locate the four corners relative to this origin to obtain their actual 3D coordinates. Given that we have both the actual 3D coordinates of the four corner points and their corresponding 2D image coordinates, we can employ the PnP algorithm to accurately estimate the pose of the corner model.

The PnP algorithm is widely used in computer vision to estimate the position and orientation of the camera relative to the object based on a set of 2D-3D points. PnP is an optimization problem that aims to determine the optimal rotation and translation matrices that minimize the reprojection error. This error is calculated by measuring the difference between the predicted positions of the 3D points when re-projected onto the image plane and their actual positions. The reprojection error is mathematically expressed as:

$$e_i = x_i - K[R \mid t]X_i \tag{5}$$

where $x_i$ represents a 2D point on the image plane, $X_i$ is the corresponding 3D point, $R$ and $t$ denote the camera's rotation matrix and translation vector, respectively, and $K$ is the camera's intrinsic parameter matrix.

The goal of PnP algorithm is to minimize the reprojection error, that is:

$$\min_{R,t} \sum_i \|e_i\|^2 \tag{6}$$

This optimization problem can be solved by iterative methods, where $R$ and $t$ are adjusted step by step through several iterations, so as to find the minimum value of the reprojection error and the optimal $R$ and $t$.

## 3 Experiments and Results

### 3.1 Dataset

We used the Intel RealSense D455 depth camera, commonly used in construction, to collect data. The D455 depth camera has two high-resolution colour and depth sensors that capture depth information through the principle of stereo vision. The intrinsic parameters of the camera are shown in Table 1 and Tabel 2 including the focal lengths $(f_x, f_y)$, the principal point coordinates $(c_x, c_y)$, and the distortion coefficients.

Table 1. Parameters of the colour camera

| Color camera | Intrinsic parameters |
|---|---|
| $f_x, f_y$ | 391.942, 391.942 |
| $c_x, c_y$ | 324.305, 239.027 |
| Distortion | [0, 0, 0, 0, 0] |

Table 2. Parameters of the depth camera

| Depth camera | Intrinsic parameters |
|---|---|
| $f_x, f_y$ | 387.781, 387.351 |
| $c_x, c_y$ | 325.001, 247.284 |
| Distortion | [-0.055, 0.066, -0.800, 8.003, -0.022] |

The coordinate systems of the depth camera and the colour camera are aligned through a transformation that includes rotation and translation. This alignment ensures that the data from both cameras are accurately matched. This transformation is defined by the extrinsic parameters of the cameras, represented by the rotation matrix $R$ and the translation vector $t$, as follows:

$$R = \begin{bmatrix} 0.99998343 & -0.00549165 & -0.00175197 \\ 0.00548666 & 0.99998093 & -0.00283836 \\ 0.00176752 & 0.00282870 & 0.99999446 \end{bmatrix}$$

$$t = \begin{bmatrix} -0.05912773311138153 \\ 0.0003899137955158949 \\ 0.00034507890813983977 \end{bmatrix}$$

We captured a total of 1018 images under different illumination conditions and camera heights, as shown in Figure 5. These 1018 images were randomly shuffled and split into two parts: 80% as the training set to train the model and 20% as the test set to evaluate the model.

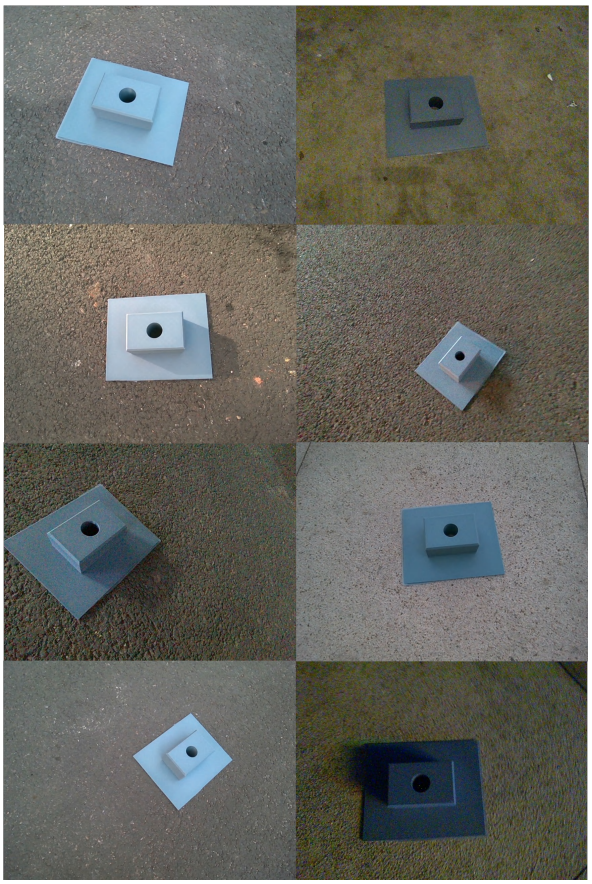

Figure 5. Image examples under different illumination conditions and camera heights.

### 3.2 Segmentation

We used the Yolov8-seg model for image segmentation, which enables instance segmentation by combining precise object detection and localization capabilities. The experiments were carried out using the NVIDIA GeForce RTX 4090 GPU. The input images were resized to a resolution of 640x640 pixels to standardize the input data format. We set the training with a batch size of 6 and the model was trained for 350 epochs.

To assess the effectiveness and accuracy of the model, we use a comprehensive set of evaluation metrics on the test set, including precision, recall, AP50 and AP50-95. These metrics help evaluate the model's ability to accurately detect and segment objects, providing a detailed analysis of its performance for segmentation.

After segmentation with Yolov8-seg, for each object, we get two key outputs: a bounding box and a segmentation mask. The bounding box marks the position of the object in the image with a rectangular box, which allows us to quickly detect the presence and approximate position of the object in the image. The segmentation mask provides more fine-grained output, including pixel-level categorization, indicating which pixels in the bounding box belong to the object and delimiting the object's boundaries.

Table 3 presents the results of the Yolov8-seg on the test set, indicating robust performance in the corner segmentation task. Furthermore, Figure 6 displays the segmentation result for the corner model using Yolov8-seg, clearly showing that precise and accurate segmentation was achieved.

Table 3. Experimental results of Yolov8-seg.

| Evaluation metrics | Box | Mask |
|---|---|---|
| Precision | 1 | 1 |
| Recall | 1 | 1 |
| AP50 | 0.995 | 0.995 |
| AP50-95 | 0.993 | 0.97 |

### 3.3 Measurement

After segmenting the surface of the corner model using the Yolov8-seg, we fit a quadrilateral to the segmented surface. The four vertices of the quadrilateral represent the

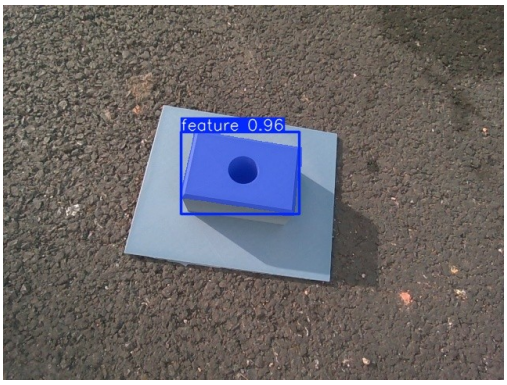


Figure 6. The segmentation result for the corner model using Yolov8-seg.

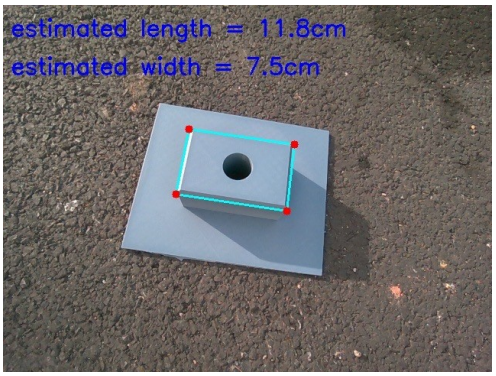


Figure 7. The measurement result for the corner model.

corners. By mapping these vertices to the corresponding depth map, we can obtain their actual three-dimensional coordinates. This enables us to measure the actual distances between the corners, thus determining the length and width of the corner model. Figure 7 shows the estimated dimensions of the corner model, with the length measured at 11.8 cm and the width at 7.5 cm. These estimations are compared to the actual dimensions, which are 12 cm in length and 8 cm in width. The test set, consisting of 204 images, was used to further validate the measurement accuracy, with the results showing that the measurement error is within 9 mm.

The measurement error can be attributed to several factors. One factor is the low resolution of the colour images, which limits the precision of segmentation and corner detection. Additionally, inaccuracies in the depth map, caused by sensor noise or resolution limitations, contribute significantly to errors in calculating the 3D coordinates of the corners. Segmentation errors from the Yolov8-seg model also play a role, as minor deviations in boundary detection can lead to slight inaccuracies in identifying the exact corner points.

### 3.4 Pose estimation

Once we have determined the exact coordinates and relative distances of the four corner points, we establish a coordinate system with the origin situated at the center of the quadrilateral formed by these points, which are at the upper left, upper right, lower left, and lower right, respectively. Following the segmentation process, we also acquire the 2D coordinates of these corner points within the colour map. Consequently, we can utilize the PnP algorithm to estimate the camera's position relative to the corner model surface. This is achieved by aligning the coordinate system of the corner model with the camera's coordinate system, allowing for precise calculation of the camera's orientation and position in relation to the corner model surface.

Figure 8 illustrates the results of estimating the pose using the PnP algorithm. The position of the camera relative to the surface of the corner model is described by a translation vector that quantifies the displacement of the camera to the model. The orientation of the camera relative to the surface of the angular model is represented by a quaternion. Quaternions provide a compact representation of orientations and do not suffer from the gimbal-lock problem that may result from orientation representations such as Euler angles.

To address occlusion, we consider introducing additional positioning tags, such as the Apriltag marker. When the corner features are not occluded, we simultaneously estimate the poses of both the Apriltag marker and the corner model, and calculate the relative pose deviation between them. When the corner model is occluded, we detect the pose of the Apriltag marker and use the previously obtained relative pose deviation between the Apriltag marker and the corner model to estimate the pose of the occluded corner model. Figure 9 shows pose estimation results in two scenarios. The upper image shows pose estimation when the corner features are not occluded, while the lower image shows pose estimation of occluded corner features using the Apriltag marker.

## 4 Conclusion

In conclusion, this study applied computer vision technology to MIC to enhance the efficiency, accuracy, and safety of the construction process. Unlike traditional construction methods, which rely heavily on manual labour and are susceptible to human error, our vision-based approach facilitates automated segmentation, measurement, and pose estimation. Specifically, we first use the Yolov8-seg model to segment the corner surfaces of modules. The

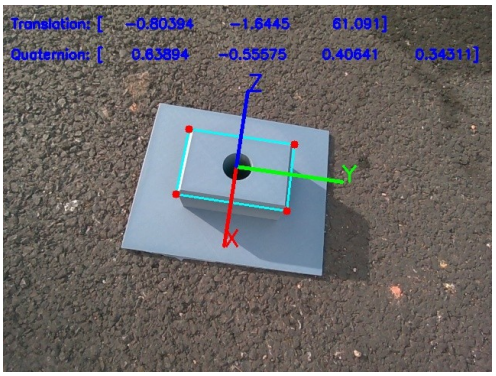


Figure 8. The pose estimation result for the corner model using PnP algorithm.

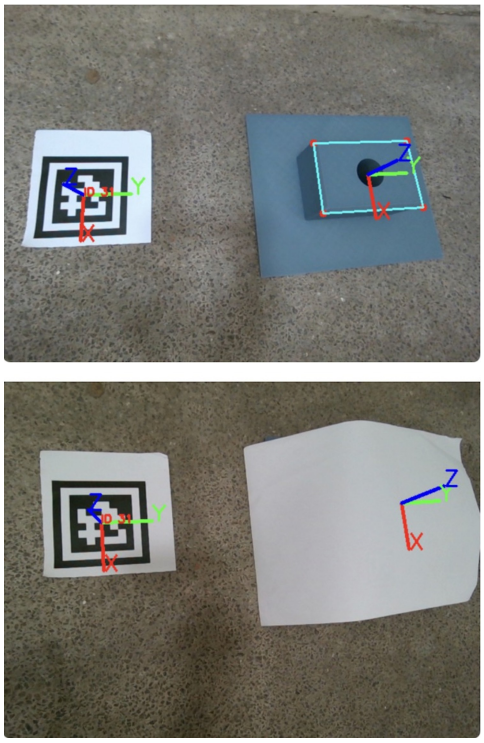

Figure 9. Pose estimation with the visible corner model (upper) and the occluded corner model using the Apriltag marker (lower).

segmented surfaces are fitted with quadrilaterals to position the four corners. By integrating the corresponding depth map, we accurately determine their 3D coordinates. Furthermore, using the 2D coordinates from the colour map and the 3D coordinates with the quadrilateral's center as the origin, we utilize the PnP algorithm for precise pose estimation. However, the 9mm measurement error is relatively high compared to the required precision in MIC. Future work will focus on algorithm optimization, addressing depth map resolution and sensor noise, and extending testing to real-world construction environments to evaluate performance under complex conditions.

## References

[1] Sherif Abdelmageed and Tarek Zayed. A study of literature in modular integrated construction-critical review and future directions. *Journal of Cleaner Production*, 277:124044, 2020. doi:10.1016/j.jclepro.2020.124044.

[2] S Marzieh Bagheri, Hosein Taghaddos, and Ulrich Hermann. Automated safety and practicality enhancement of lift plans in modular construction. *Automation in Construction*, 168:105731, 2024. doi:10.1016/j.autcon.2024.105731.

[3] Kamyab Aghajamali, Ala Nekouvaght Tak, Hosein Taghaddos, Ali Mousaei, Saeed Behzadipour, and Ulrich Hermann. Planning of mobile crane walking operations in congested industrial construction sites. *Journal of Construction Engineering and Management*, 149(7):04023047, 2023. doi:10.1061/JCEMD4.COENG-13109.

[4] Shuyuan Xu, Jun Wang, Wenchi Shou, Tuan Ngo, Abdul-Manan Sadick, and Xiangyu Wang. Computer vision techniques in construction: a critical review. *Archives of Computational Methods in Engineering*, 28:3383–3397, 2021. doi:10.1007/s11831-020-09504-3.

[5] Siddharth Singh Chouhan, Ajay Kaul, and Uday Pratap Singh. Image segmentation using computational intelligence techniques. *Archives of Computational Methods in Engineering*, 26:533–596, 2019. doi:10.1007/s11831-018-9257-4.

[6] Farhana Sultana, Abu Sufian, and Paramartha Dutta. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowledge-Based Systems*, 201:106062, 2020. doi:10.1016/j.knosys.2020.106062.

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. doi:10.1109/CVPR.2015.7298965.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*,

pages 234–241. Springer, 2015. doi:10.1007/978-3-319-24574-4_28.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. doi:10.1109/TPAMI.2017.2699184.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. doi:10.1109/ICCV.2017.322.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. doi:10.1109/TPAMI.2016.2577031.

[12] Leanne Attard, Carl James Debono, Gianluca Valentino, Mario Di Castro, Alessandro Masi, and Luigi Scibile. Automatic crack detection using mask r-cnn. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 152–157, 2019. doi:10.1109/ISPA.2019.8868619.

[13] S Dilaksha, KATO Ranadewa, D Weerasooriya, Agana Parameswaran, and Panchali Weerakoon. Comparative analysis of challenges in manual and automated construction progress monitoring in sri lanka. In *Proceedings The 12th World Construction Symposium— August*, page 380, 2024. doi:10.31705/WCS.2024.30.

[14] Randolph Blake and Hugh Wilson. Binocular vision. *Vision research*, 51(7):754–770, 2011. doi:10.1016/j.visres.2010.10.009.

[15] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016. doi:10.1007/s00138-016-0784-4.

[16] Song Zhang. High-speed 3d shape measurement with structured light methods: A review. *Optics and lasers in engineering*, 106:119–131, 2018. doi:10.1016/j.optlaseng.2018.02.017.

[17] Bing Liu. Long-distance recognition of crack width in building wall based on binocular vision. In *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture*, page 1908–1912, New York, NY, USA, 2022. doi:10.1145/3495018.3495513.

[18] Haotian Yu, Yu Huang, Dongliang Zheng, Lianfa Bai, and Jing Han. Three-dimensional shape measurement technique for large-scale objects based on line structured light combined with industrial robot. *Optik*, 202:163656, 2020. doi:10.1016/j.ijleo.2019.163656.

[19] Xiao Xin Lu. A review of solutions for perspective-n-point problem in camera pose estimation. In *Journal of Physics: Conference Series*, volume 1087, page 052009. IOP Publishing, 2018. doi:10.1088/1742-6596/1087/5/052009.

[20] Chao Mi, Yi Liu, Yujie Zhang, Jiaqi Wang, Yufei Feng, and Zhiwei Zhang. A vision-based displacement measurement system for foundation pit. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023. doi:10.1109/TIM.2023.3311069.

[21] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. doi:10.1109/ICCV.2015.336.

[22] Prabesh Paudel and Kyoung-Ho Choi. A deep-learning based worker's pose estimation. In *Frontiers of Computer Vision: 26th International Workshop, IW-FCV 2020, Ibusuki, Kagoshima, Japan, February 20–22, 2020, Revised Selected Papers 26*, pages 122–135. Springer, 2020. doi:10.1007/978-981-15-4818-5_10.

[23] Roshan Panahi, Joseph Louis, Ankur Podder, Colby Swanson, and Shanti Pless. Automated assembly progress monitoring in modular construction factories using computer vision-based instance segmentation. In *Computing in Civil Engineering 2023*, pages 290–297. 2024. doi:10.1061/9780784485224.036.

[24] Zhenjie Zheng, Zhiqian Zhang, and Wei Pan. Virtual prototyping-and transfer learning-enabled module detection for modular integrated construction. *Automation in Construction*, 120:103387, 2020. doi:10.1016/j.autcon.2020.103387.

[25] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4): 1680–1716, 2023. doi:10.3390/make5040083.