Enhanced YOLOv8 Algorithm for Real-Time Instance Segmentation of Components in Modular Integrated Construction

Xinqi Liu[1*] and Wei Pan[2]

[1]Ph.D candidate, Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
[2]Professor, Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
Corresponding author's E-mail: Xinqi_liu28@connect.hku.hk

Abstract

Detecting module components at the factory is crucial for safety monitoring, quality control, and productivity enhancement. However, traditional segmentation methods are neither cost-effective nor capable of achieving real-time performance. To address these challenges, this study proposes an improved YOLOv8 modular integrated construction segmentation algorithm. The proposed method introduces the construction of a small object-YOLO, optimizing the YOLOv8 model by replacing the basic module with a novel cross-stage partial network fusion module. This new module employs deformable convolutional networks v2 to manage geometric variations of objects and focus on relevant image regions. Additionally, the Wise-IoU strategy reduces the competitiveness of high-quality anchor boxes and mitigates harmful gradients generated by low-quality examples. The Multi-Head self-attention mechanism further enhances detection accuracy by capturing the relationship between the image and significant objects, making it more suitable for the modular integrated construction dataset. Given that construction images are often taken from a top or bird's-eye view, small objects can be challenging to be detected. Therefore, this algorithm incorporates a small object detection algorithm to improve the model's capability in identifying small objects. Experimental results demonstrate that the improved YOLOv8 model effectively identifies moving objects, achieving a 4.4% increase in mAP and a 4.3% increase in F1 score compared to the original YOLOv8 model, while reducing parameters by 54.05% and GFLOPs by 55.39%. The proposed algorithm provides a reference for automatic segmentation methods of modular integrated construction components at the factory.

Keywords: YOLOv8, segmentation, Construction factory, Modular integrated construction

1. INTRODUCTION

The utilization of computer vision techniques for automated quality management has garnered significant attention from both academia and industry. These techniques leverage digital videos and images obtained from modular integrated construction (MiC) factories through cost-effective tools such as digital cameras, surveillance cameras, smartphones, and unmanned aerial vehicles (UAVs). . MiC builds on the modular construction approach by integrating advanced production technologies into the re-engineered building design and construction process. It employs three-dimensional (3D) units that are fully factory-finished internally and transported to sites for installation. Defined by Pan and Hon (Pan and Hon, 2020) as a "game-changing disruptively-innovative approach," MiC transforms fragmented site-based construction into an integrated, value-driven production and assembly of prefinished modules, enhancing quality, productivity, safety, and sustainability. Various applications have emerged, including productivity analysis, progress monitoring (Kopsida, Brilakis and Vela, 2015), and safety supervision (He *et al.*, 2020). These applications depend on the precise detection, segmentation, and accurate identification of mobile entities (e.g., beams, columns, and slabs) within MiC factories, representing a fundamental necessity (Rebolj *et al.*, 2008). However, previous studies predominantly employed detectors based on manual features, significantly limiting the generalization of applications.

The detectors used in these studies were trained and evaluated using specialized datasets, which are often limited in scope and customized (e.g., tailored for specific objects or sourced from restricted sites, devices, viewpoints, or under varying weather and lighting conditions). Consequently, detectors trained on such confined data sources may exhibit significant variability and struggle to generalize effectively to the MiC factory environment (Kim *et al.*, 2019). Therefore, it is imperative to develop a comprehensive detector model imbued with diversity, specifically designed for module production, and made openly accessible to the public. In pursuit of this objective, 1,124 images were collected, labeling 3,234 components in collaboration with three MiC manufacturers located in Guangdong Province of China. This comprehensive dataset, referred to as "Modular Integrated Construction Segmentation" (MiC-seg), was utilized to train the model for the MiC factory.

Developing a suitable detector model for MiC production presents several challenges. Prior research on object detection primarily falls into two categories: The first category encompasses region-based two-stage detection models characterized by a bifurcated process. The initial step involves proposing multiple regions that might contain objects. In the second step, a classification network is employed on these proposed regions to ascertain the object category within each delineated area. Prominent algorithms within this two-stage paradigm include fast region-based Convolutional Neural Networks (Fast R-CNNs) (Girshick, 2015), Region-Based Fully Convolutional Networks (R-FCNs) (Dai *et al.*, 2016), and mask-region-based Convolutional Neural Networks (Mask R-CNNs) (He *et al.*, 2017). The second category involves a one-stage detection approach grounded in regression. This methodology directly distinguishes specific categories and performs regression on their boundaries.

2. LITERATRUE REVIEW

Previous research on the identification of construction objects and materials has primarily relied on manually designed features extracted from digital images to recognize specific visual elements. Zou and Kim (2007) employed the hue, saturation, and value (HSV) color space to identify excavators in construction site images. A threshold value for saturation was used to distinguish relatively colorful objects (e.g., excavators) from achromatic backgrounds (e.g., dark soil or white snow). Huo et al. (2020) proposed an automatic method for detecting engineering structural components based on the Deeply Supervised Object Detector (DSOD) algorithm. This approach was validated using a comprehensive image dataset of engineering structural components, acquired through multilayer, polymorphic, multidirectional, and multi-angle data collection. Zhu et al. (2010) applied Canny edge detection (Canny, 1986) and Hough transform techniques (Mukhopadhyay and Chaudhuri, 2015) to detect column edges, followed by an object reconstruction method to locate and quantify the number of columns present in an image or video. Kim et al. (2016) utilized a scene-parsing technique (Liu, Yuen and Torralba, 2011) to identify distinct construction objects in a query image by matching them with labeled images from a database and transferring the labels of the best-matching candidate images to the query image. One of the most widely used and advanced algorithms for object detection is the Region-based CNN, also known as R-CNN, introduced by (Girshick et al., 2014). R-CNN employs a selective search to detect regions of interest (RoIs), then utilizes a CNN to extract features from each region, and ultimately applies an SVM for object classification within those regions (Vaidya et al., 2023). However, faster variants such as Fast R-CNN have been introduced to address the excessive time and space requirements of running this algorithm. Ren et al. (2015) described a Faster R-CNN as being composed of two components: a fully convolutional network called the Region Proposal Network (RPN), which suggests regions of interest (ROIs). YOLO (Redmon et al., 2016) and single-shot multi-box detector (SSD) algorithms differ from region proposal-based methods in that they combine classification and localization tasks into a single neural network, resulting in a significant reduction in computational burden.

3. METHODOLOGY

A hybrid research strategy combining qualitative and quantitative methods was employed to develop an instance segmentation algorithm for MiC factories. First, a comprehensive literature review was conducted to examine the challenges of high-rise MiC plant production. It was identified that there is no existing automatic identification model for MiC plant production, highlighting the significance of developing such a model for automated MiC production. Secondly, our proprietary dataset was used for annotation and combined training, specifically tailoring the model for MiC components. Thirdly, the algorithm evaluation system was utilized to assess the system's performance. The detailed methodology for developing and evaluating the system is described in the following sections.

3.1. Optimized YOLOv8

According to the characteristics of the MiC component, this paper proposed MiC Component-YOLO (MiCC-YOLOv8). Although YOLOv8 was declared the highest-performing YOLO model ever released, it still falls short of our detection requirements. In a MiC module factory, many small objects are located far from the top view, which makes it challenging for YOLOv8 to capture the necessary feature information for these small objects because of YOLOv8's deeper feature maps with relatively large down-sampling multiples. As shown in Figure 1 (a), a new framework has been proposed to address this challenge. This framework accelerates YOLOv8's ability to detect small objects in real time in factory environments. MiCC-YOLO is an optimized and improved version of YOLOv8 that works by decomposing an image into grid cells that predict B-bounding boxes.

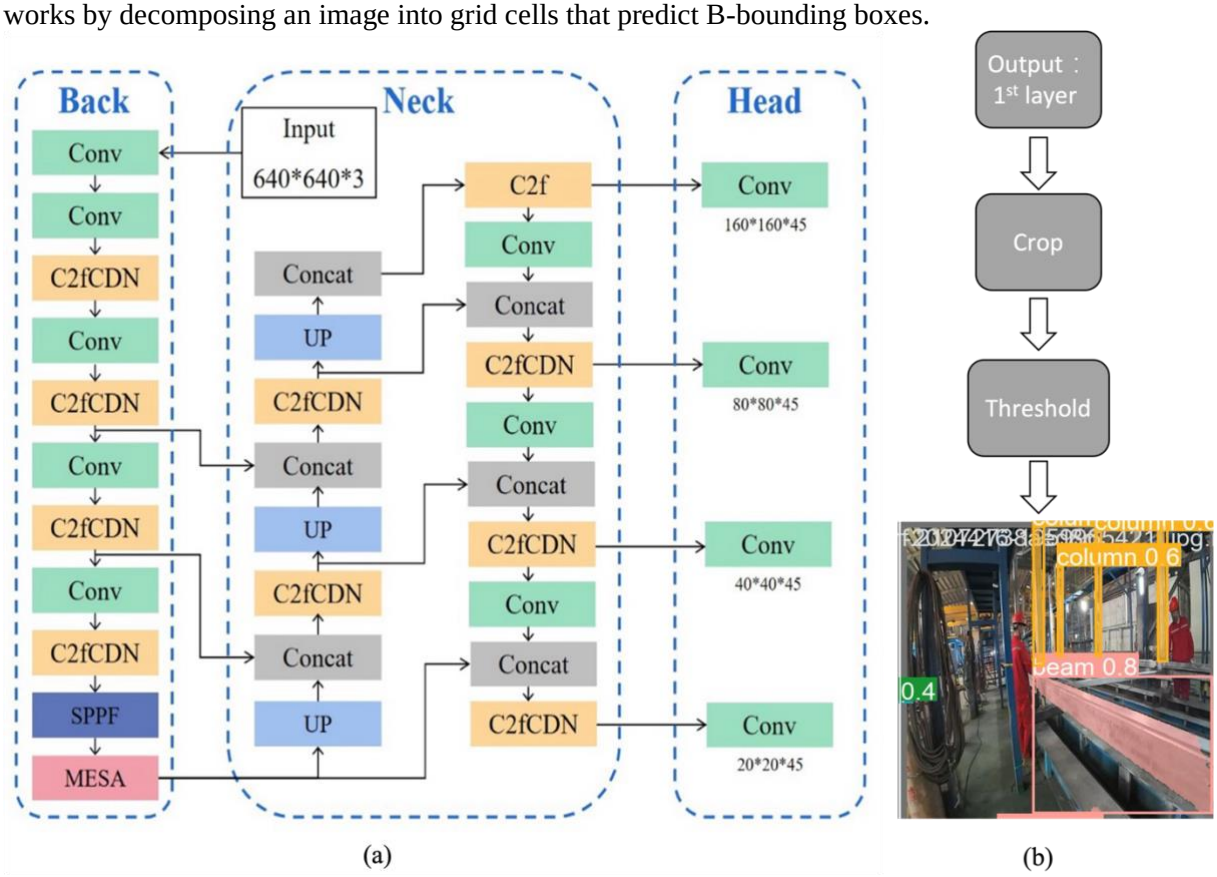

Figure 1 (a)Architecture of MiC component-YOLO (b)Instance segmentation prediction process.

The Bounding boxes are designed to regress the location and predict the confidence level, representing the confidence that the predicted box contains an object and the accuracy of box prediction.

As shown in Figure 1 (b), the mask of the target instance was cropped to zero at a location outside the instance box. The ground truth detection box was used for training, and the detection box obtained from the object detection was used for testing.

3.2. Multi-head self-attention

YOLO is a popular real-time object detection algorithm that has gained significant attention because of its impressive accuracy and speed. However, YOLO faces challenges detecting small objects or objects with complex backgrounds. To address these challenges, attention mechanisms have been integrated into the YOLO architecture, leading to improved performance in object detection tasks. Attention mechanisms allow the model to focus on certain regions of an image, which is particularly useful for detecting objects in cluttered scenes.

Multi-head self-attention (MHSA) is an approach that has proven effective in Neural News Recommendation (Wu et al., 2019). The attention mechanism is a technique that can be widely used in computer vision, natural language generation, and recommendation systems.

3.3. Small objects detection

In this study, small targets were recognized through the astute deployment of an oversampling strategy. This strategy orchestrates the augmentation of small target prominence within the loss function, thereby improving the performance of small target detection. During the training phase, oversampling was judiciously applied to the images of small targets, which is a strategy devised to alleviate the scarcity of such instances. This was achieved by replicating and oversampling samples harboring small targets. The extent of replication, quantified as the oversampling rate, was meticulously regulated to equalize the representation of large and small samples, thereby mitigating bias.

3.4. Deformable convolution V2 for C2f

As the target is dynamic in construction scenarios, geometric variations caused by scale, pose, viewpoint, and part deformation pose significant challenges to object recognition and detection. To address this issue, this study introduces a modified version of Deformable ConvNets CSPDarknet53 to 2-Stage FPN (C2fDCN), which enhances its capability to concentrate on relevant image regions. This enhancement was achieved through improved modeling power and more robust training techniques. The outstanding performance of deformable Convolutional Networks stems from their capacity to adjust to the geometric variations in objects. By analyzing its adaptive behavior, it is evident that the spatial support for its neural features aligns more closely with the object structure than with regular ConvNets. However, this support might extend beyond the region of interest, leading to features being influenced by irrelevant image content.

4. RESULTS AND ANALYSES

One of the reasons YOLO struggles with small target detection is the small size of the target samples coupled with the relatively large downsampling multiplier of YOLOv8. This makes it difficult for deeper feature maps to obtain the parameters based on the output of the previous layer. For the MHSA, the width and height were set to 14 with four heads, and the number of channels was the same as that of the output of the previous layer. For example, e_lambda (Wu et al., 2019) was set to 1e-4. As shown in Table 1, the parameter-dependent attention mechanisms performed significantly better than SimAM without parameters. Therefore, for YOLO optimization, using a parameter-free model such as SimAM is not recommended. The mAP of MHSA was significantly better than that of shufflattention for all classes. Additionally, F1, a combined comparison of precision and recall, was

better for the MHSA. Consequently, the final MiCC-YOLO model was constructed using the MHSA model. Notably, this improvement in the model performance was not only significant for the building scenario but also crucial for YOLO optimization. The final MiCC-YOLO model improved by 0.028 (4%) for all classes of mAP@0.5 and 0.038 (5%) for the Best. Thus, these findings are of great importance for YOLO optimization and the module factory.

Table 1 Comparison of the detection results of different algorithms.

Model	**Tests1-F1(%)**	**Tests2-F1(%)**
YOLOv8	66.21	66.13
Ensemble (YOLO-v4+Faster-RCNN)	56.36	57.07
EfficientDet	56.5	54.7
YOLOv4	55.4	54.1
YOLO model trained on CSPDarknet53 backbone	58.14	57.51
Multi-stage Faster R-CNN with Resnet-50 and Resnet-101 backbones	53.68	54.26
Road Damage Detector using Detectron2 and Faster R-CNN	51	51.4
FR-CNN;Classifying the region and using regional experts for the detection	47.20	46.56
Ours	72.15	72.23

As shown in Table 1, the proposed algorithm achieved promising results on the MiC component segmentation task verification set. They obtained an F1 score of 72.15% and an mAP value of 74.3%. These results indicate an improvement over the initial YOLOv8 algorithm, with an increase of 8.9% in the F1 score and 4.35% in mAP. For the test sets, the algorithm achieved an F1 score of 35.82% and a mAP score of 39.05%. Furthermore, the algorithm exhibited efficiency in terms of its parameters and number of calculations, with GFLOPs of 35.2 and 11.9 M parameters. Compared to the original YOLOv8 algorithm, which has GFLOPs of 28.6 and 11.2 M parameters, the proposed algorithm demonstrates superior performance. The algorithm combines multi-head self-attention with a deformable convolution module, incorporates the small-object algorithm, and utilizes Wise-IoU. The effectiveness of the algorithm is confirmed experimentally, and the detection effect is illustrated. Figure 2 illustrates the visual segmentation method used to process the sample image. Through the model, images can be segmented, with walls, beams, and columns clearly identified and separated.

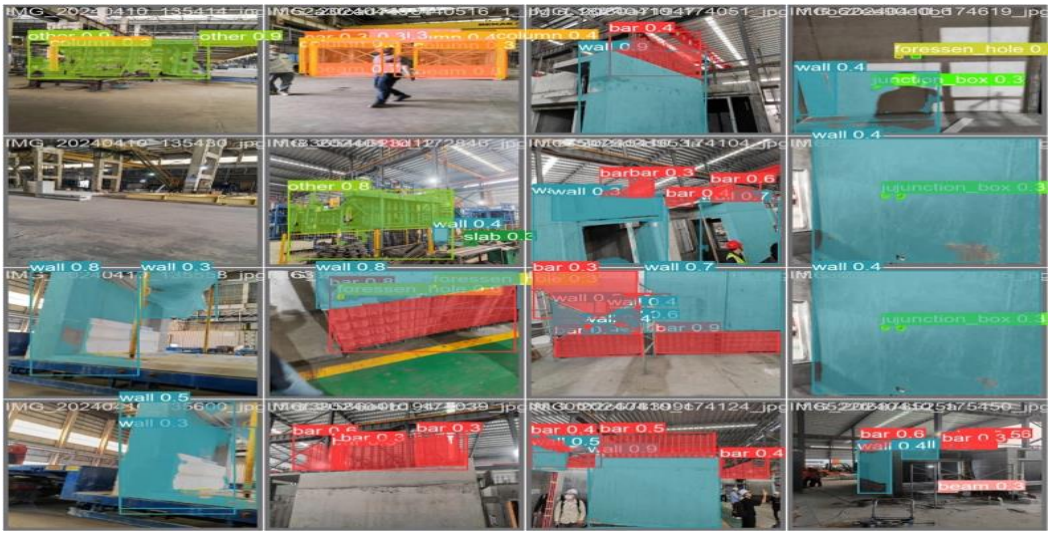

Figure 2 The result of segmentation module

5. CONCLUSIONS

The component instance segmentation process at the MiC factory, which leverages computer vision,

forms the bedrock of a smart manufacturing infrastructure crucial for automating production. As smart manufacturing evolves, challenges have emerged in accurately detecting and segmenting MiC components in real-world scenarios and images. To tackle this issue, this paper develops the MiCC-YOLOv8 algorithm, which involves an attention mechanism, a small-object detection system, and an optimized loss function.

The developed method initially employs a multi-head self-attention system to enhance the network model for detecting diverse module components. It was observed that the parameterized attention mechanism improves YOLO's mAP efficiency compared to its parameter-free counterpart. In complex construction environments, incorporating a deformable convolution module alongside the YOLOv8 algorithm significantly improves object detection accuracy. Experimental findings demonstrate that our model achieves a mAP of 0.74, surpassing YOLOv8's performance of 0.712. This superior capability positions our novel method as ideal for accuracy-dependent scenarios with constraints on memory and computing power, such as embedded device functions. Furthermore, the integration of a small-object detection layer significantly boosts accuracy, particularly beneficial for quantity segmentation and safety applications during module production, especially when utilizing wide-angle cameras.

However, achieving a balance between real-time performance and algorithmic accuracy necessitates an increase in model parameters compared to the original version. Future enhancements should focus on optimizing the algorithm to reduce the parameter count, making it more suitable for application in manufacturing settings with limited computational resources.

6. ACKNOWLEDGMENTS

The work described in this paper was supported by the Research Impact Fund of the Hong Kong Research Grants Council (Project No.: HKU R7027-18).

7. REFERENCES

Canny, J. (1986) 'A computational approach to edge detection', *IEEE Transactions on pattern analysis and machine intelligence*, (6), pp. 679–698.

Dai, J. *et al.* (2016) 'R-fcn: Object detection via region-based fully convolutional networks', *Advances in neural information processing systems*.

Girshick, R. (2015) 'Fast r-cnn', in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

He, C. *et al.* (2020) 'Effects of Safety Climate and Safety Behavior on Safety Outcomes between Supervisors and Construction Workers', *Journal of Construction Engineering and Management*, 146(1), p. 04019092.

Kopsida, M., Brilakis, I. and Vela, P.A. (2015) 'A review of automated construction progress monitoring and inspection methods', in *Proc. of the 32nd CIB W78 Conference 2015*, pp. 421–431.

Mukhopadhyay, P. and Chaudhuri, B.B. (2015) 'A survey of Hough Transform', *Pattern Recognition*, 48(3), pp. 993–1010.

Pan, W. and Hon, C.K. (2020) 'Briefing: Modular integrated construction for high-rise buildings', *Proceedings of the Institution of Civil Engineers - Municipal Engineer*, 173(2), pp. 64–68.

Rebolj, D. *et al.* (2008) 'Automated construction activity monitoring system', *Advanced engineering informatics*, 22(4), pp. 493–503.