IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 24 October 2023, accepted 29 November 2023, date of publication 7 December 2023,
date of current version 14 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340895

RESEARCH ARTICLE

Automated Construction Site Monitoring Based on Improved YOLOv8-seg Instance Segmentation Algorithm

RUIHAN BAI[1], MINGKANG WANG[2], ZHIPING ZHANG[2], JIAHUI LU[2], AND FENG SHEN[3]

[1]School of Civil and Transportation Engineering, Hohai University, Nanjing, Jiangsu 210098, China
[2]School of Civil Engineering, Tongji University, Shanghai 200000, China
[3]School of Civil Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu 215011, China

Corresponding author: Feng Shen (shenfeng1023@163.com)

This work was supported in part by the Project on the Integration of Industry, Education and Research of Jiangsu Province under Grant BY2022570; in part by the Qinglan Project of Jiangsu Province, in 2022; and in part by the IoT Technology Application Transportation Industry Research and Development Center, Hangzhou, in February 2023.

ABSTRACT Utilizing Unmanned Aerial Vehicles (UAV) and instance segmentation for construction site monitoring(such as construction machinery and operation surfaces) offers a significant leap in management efficiency over traditional manual supervision methods. However, in UAV-based remote sensing images, the subtle presence of construction machinery and the image features resemblances among various operational surfaces make it difficult to segment instances. To address these challenges, this study proposed a novel instance segmentation model based on the YOLOv8-seg model. Given the unique challenges, the proposed model makes three improvements to the original YOLOv8-seg model. First, the paper incorporates the FocalNext module, which extends the sense field of the convolutional kernel to capture contextual data and integrates multilevel features, enhancing the perception of local details. Second, the paper incorporates the Efficient Multiscale Attention (EMA) module, which refines image features by emphasizing spatial-channel interactions and adeptly contrasts patterns across scales to detect nuances overlooked by conventional models, aiding in distinguishing similar construction operation surfaces. Last, given the intricate nature of construction site images, this paper incorporates the Context Aggregation module, which enhances pixel analysis by intelligently modulating feature weights to highlight essential global contexts. The ablation experiment demonstrates that the enhancements perform well on the YOLOv8-seg two variants model. Comparative experimental results show that the improved model significantly outperforms existing instance segmentation models regarding model performance, complexity, and inference speed. Overall, the improved YOLOv8-seg model balances model performance and computational complexity to meet the needs of edge device deployment in field monitoring.

INDEX TERMS Instance segmentation, YOLOv8, construction site.

I. INTRODUCTION

In the construction industry, the progress of construction work surfaces and the working conditions of machinery are two crucial factors. Specifically, construction progress represents not only the real-time completion of the project but also a direct reflection of the optimization of the project's economy, time, and resources. Machinery's working condition directly affects the construction efficiency and quality. Any lag in progress or machinery failure may trigger cost increases, project quality decline, and even safety accidents. Effective supervision ensures that projects are carried out according to the intended schedule and budget while safeguarding the quality and safety of the work. However, traditional monitoring relies heavily on manual observation, recording, and reporting, which has multiple

The associate editor coordinating the review of this manuscript and approving it for publication was Halil Ersin Soken.

limitations. Firstly, manual recording may lead to omission or bias of information, which is the key basis for project decision-making. Second, frequent manual inspections are labor-intensive and difficult to achieve real-time, continuous monitoring. With the development of digitalization and automation in the construction industry, there is an increasing demand for more advanced, accurate, and efficient monitoring methods. Modern monitoring methods can capture construction progress and machinery status information in real-time and automatically, providing timely and accurate data support for decision-makers, thus optimizing resource allocation, improving construction efficiency, and guaranteeing project quality.

In recent years, the field of computer vision has seen significant advancements, notably in the development of instance segmentation techniques. Unlike traditional image analysis methods, instance segmentation offers a detailed recognition of object instances down to the pixel level, boasting enhanced resolution and accuracy. Several advanced algorithms based on convolutional neural networks have emerged in recent years, such as Mask R-CNN [1], PANet [2], PointRend [3], and SOLO [4]. Among others, Mask R-CNN builds upon Faster R-CNN [5] by adding the capability to predict segmentation masks, while the RoIAlign module addresses alignment discrepancies. PANet further amplifies the flow of feature information, offering robust support for smaller instances. PointRend treats the segmentation problem as a rendering task, producing high-quality details around object boundaries. In contrast, SOLO employs a direct prediction approach, eliminating the need for RoI to present a streamlined and efficient solution. The YOLO (You Only Look Once) series, primarily known for its real-time object detection capabilities, has also inspired methods incorporating instance segmentation features. YOLACT [6] is a real-time instance segmentation method that builds upon the YOLO framework. Instead of predicting raw masks for every instance, it predicts a set of prototype masks for the whole image and linear combination coefficients for each detected instance. As an adaptation of the YOLOv5 model, YOLOv5-seg [7] extends its architecture to handle pixel-wise mask and bounding box predictions. Similarly, building on the YOLOv7 object detection framework, YOLOv7-seg [8] is devised to handle instance segmentation.

The practical implications of instance segmentation algorithms are vast. In medicine, it offers the possibility to locate and measure structures in the body precisely, thus optimizing the diagnostic and therapeutic process; in agriculture, it can provide more powerful data support for agricultural production and pest control by analyzing farmland images in detail; in automated driving and robotics, instance segmentation provides higher accuracy for environment sensing and object interaction. Likewise, numerous research studies have integrated computer vision into construction site management to enhance construction site operations [9], [10], [11], [12], [13]. Yu and Nishio [9] propose a computer vision-based instance segmentation framework for multilevel bridge inspection, focusing on structural component detection and segmentation. Xiao et al. [10] propose a vision-based method for tracking workers in off-site construction by integrating deep learning instance segmentation. The Mask R-CNN algorithm is used for instance segmentation, and a matrix-based association approach is employed for tracking. Kang et al. [11] propose a deep learning-based one-stage instance segmentation model for surveillance camera systems at construction sites, considering weather conditions such as rain, snow, and fog. Kumar et al. [12] propose a Mask R-CNN-based approach for automatic multiclass instance segmentation of concrete damage. Fang et al. [13] propose a sewer defect detection framework for sewer floating capsule robots. The framework includes instance segmentation, real-time localization, and 3D reconstruction.

In recent years, the rapid development of unmanned aerial vehicle (UAV) technology has brought about revolutionary changes in various industries. UAV have emerged as crucial inspection instruments, offering a more streamlined and precise data gathering and analysis method. As a data acquisition platform and measurement instrument, UAV systems are becoming attractive for many surveying applications [14]. Unlike traditional methods, UAV-captured remote sensing images provide unrestricted views, ensuring holistic observations from any perspective or location. Consequently, the development of instance segmentation technology tailored for UAVs broadens the horizons of their potential applications. Specifically, Xie et al. [15] propose a method for tree crown extraction in high canopy-density forests using UAV remote sensing images and instance segmentation models. Song [16] researched a vehicle instance segmentation algorithm based on UAV aerial images for traffic monitoring systems. Wang [17] focuses on the research of extracting traffic signs from tilted UAV images using the Mask R-CNN instance segmentation framework. The study develops a method to address the challenges of perspective deformation, large-scale variation, and occlusion in tilted UAV images. Stewart et al. [18] propose a deep learning-based approach using Mask R-CNN for quantitative Northern Leaf Blight (NLB) quantitative phenotyping in UAV images. The study demonstrates the potential of combining UAV technology and deep learning for high-throughput and accurate quantitative measurement of plant diseases. Weyler et al. [19] propose a vision-based approach for joint instance segmentation of crop plants and leaves in agricultural fields and breeding plots using UAV imagery. Liu and Chou [20] propose a Bayesian-optimized deep-learning model for UAV images to identify and segment deterioration patterns underneath bridge decks.

Monitoring construction progress and quality is essential in current architectural construction management. However, traditional manual inspection methods are inefficient and may be constrained by various objective conditions. Thus, this paper proposes an instance segmentation model based on engineering images collected by UAV for construction machinery and operating surface instance segmentation. The ability of UAV to frequently fly over the site allows for capturing substantial real-time remote sensing images. This allows construction managers to perform ongoing and immediate site monitoring, considerably enhancing management efficiency. Instance segmentation can provide high-precision semantic and location analysis to support civil engineering managers in more efficient supervision and planning. The instance segmentation model can be employed to edge devices (UAV) to monitor safety and quality issues during construction processes, promptly identifying and addressing potential risks. By obtaining real-time information on construction machinery and operation surfaces, such as objects' shapes, categories, and positions, engineering managers can better understand and evaluate on-site conditions, facilitating more informed decision-making. However, computer vision-based construction scenario instance segmentation is uniquely challenging. First, construction machinery typically occupies only a tiny portion of the image in UAV-based remote sensing images. The condition causes the characteristics of small objects to be overshadowed by extensive background information during image recognition and segmentation, posing challenges for identifying small targets. Moreover, there are often many operational surfaces with similar properties in construction sites, such as concrete surfaces and gravel floors, which may exhibit remarkably similar visual characteristics in remote sensing images captured by UAV, increasing the difficulty of identification and segmentation. Finally, real-time monitoring for the management of construction sites is required, and the accumulation of delays may lead to distortion and misjudgment of results. The solution is to use edge computing, integrating deep learning processing units on UAV and performing the computational tasks near the data source, thus reducing data transmission and processing latency. Deep learning algorithms are computationally complex, while lightweight models often perform poorly. Thus, proposing a highly accurate algorithm that is lightweight is crucial for the edge deployment of instance segmentation algorithms.

Over the past few years, many researchers have favored the YOLO series for its real-time object detection capabilities. As the state-of-the-art algorithm of the YOLO series, The YOLOv8-seg model has the following potential advantages over mainstream instance segmentation models: 1. YOLO series algorithms emphasize real-time target detection and segmentation. The YOLOv8-seg inherits this feature and can provide near real-time instance segmentation, making it ideal for applications that require immediate feedback, such as real-time surveillance at construction sites. 2. The YOLO series serves as a single-stage detector, eliminating the need for a complex two-stage process or area proposal mechanism. This design makes the model simpler and more efficient for actual deployment. 3. The YOLO series is known for its simplicity of training, requiring fewer resources and less time than some other deep learning models. This is especially beneficial for applications that require retraining on specific construction site data. 4. As a state-of-the-art model, YOLOv8-seg can incorporate the latest computer vision and deep learning advances. This ensures it stays on the cutting edge, outperforming older models in various scenarios. Based on the advantages analyzed above, this study has selected YOLOv8-seg as the primary segmentation model. Given the unique challenges associated with instance targeting, we have improved the YOLOv8-seg model to develop an instance segmentation model specifically for construction sites. The research improvement of the model can be summarised as follows:

- The paper introduces the FocalNext module, specifically designed to address the challenges of small object segmentation in remote sensing images. By broadening the receptive field of the convolutional kernel, the module captures more contextual information, which is crucial for distinguishing small objects from their surroundings. Additionally, integrating multi-level image features ensures detailed recognition of these objects. Together, these enhancements facilitate the effective parallel processing of fine local details and the broader scene, resulting in a marked improvement in small object segmentation accuracy.
- The paper introduces the Efficient Multi-scale Attention (EMA) module to distinguish visually similar construction operation surfaces. The module refines image features by emphasizing the interaction between spatial and channel dimensions. By effectively comparing patterns at different scales, it can identify subtle dependencies ignored by original models.
- Considering the complexity of construction site imagery, this paper introduces the Context Aggregation module. The module refines pixel analysis by smartly adjusting feature weights, emphasizing key global contexts while reducing irrelevant information.

The remainder of this paper is organized as follows. Section II presents details about the model improvements. Section III describes the data used in this study, evaluation metrics, and experimental environment configuration. Section IV presents the analysis of the experimental results. Section V is the conclusion of the paper. Finally, Section VI presents the limitations and future work of this paper.

II. METHODS

A. FUNDAMENTALS OF THE YOLOv8-seg MODEL

YOLOv8, released by Ultralytics, outperforms its YOLO antecedents in terms of speed and accuracy. Overall, the backbone of the YOLOv8 model references the design concept of YOLOv7 ELAN: the basic convolutional module is the C2f module, which integrates two parallel gradient flow branches, facilitating a more robust gradient information flow. Additionally, YOLOv8 uses the Spatial Pyramid Pooling Fusion (SPPF), a module to extract contextual information from images at varying scales that significantly enhances the model's generalization capabilities. In YOLOv8's neck design, the model removed convolutional structures during

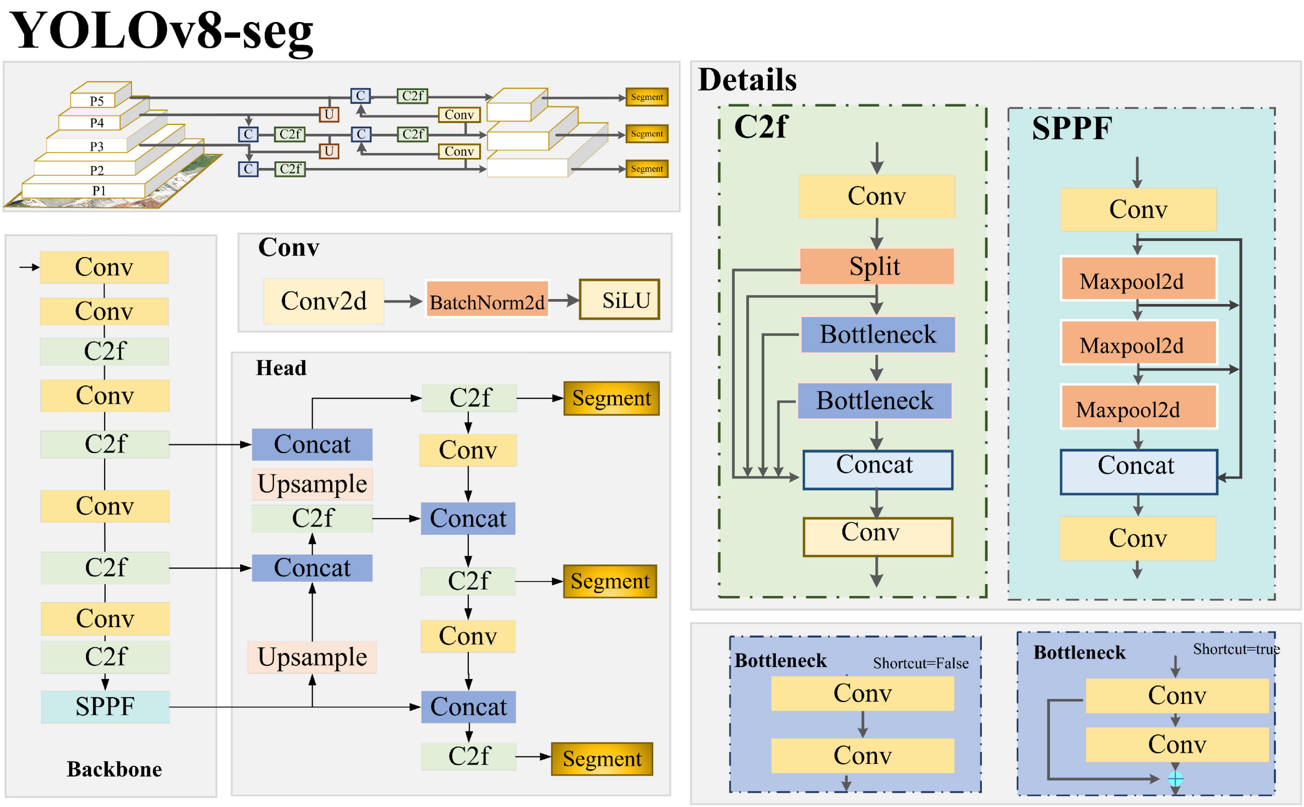


FIGURE 1. The model structure of the YOLOv8-seg model.

the up-sampling phase and strategically replaced the C3 module with the C2f module. YOLOv8 provides a framework for model training, enabling the performance of essential tasks such as object detection, instance segmentation, image classification, and pose estimation. Among them, the model structure of the YOLOv8-seg model is shown in **Fig. 1**.

B. IMPROVED YOLOv8-seg MODEL

1) FOCALNEXT BLOCK

In remote sensing image analysis, accurately segmenting small objects is challenging. The challenges primarily stem from the targets' small scale and the background's complexity. To address these issues, this paper introduces a novel computer vision module, FocalNext block [21], which combines the strengths of dilated depthwise convolutions and skip connections to simultaneously process fine-grained local information and coarse-grained global information, thereby enhancing small object segmentation performance. As depicted in **Fig. 2A** and **Fig. 2B**, the FocalNext block is an augmented variant of the ConvNext block [22], endowed with an additional dilated depth-wise convolution and two skip connections:

- Dilated depthwise convolutions, a pivotal component of the FocalNext block, expand the receptive field of the convolutional kernel, enhancing the model's capacity to perceive global information. Given that the small targets often have a considerably small scale compared to their surrounding environment, the segmentation model with a large enough receptive field can effectively capture a broader range of background information, thus increasing the understanding of information about the environment around the small target.
- Given the considerable variations in the size and distribution of ground features, small targets may seem substantially diminutive compared to their surroundings. Global and local information complement each other and serve as a key strategy in addressing the problems in remote sensing imagery. Skip connections enable the model to fuse features from varying imagery feature levels, enhancing the model's ability to utilize fine-grained local information for accurate small target detection while leveraging coarse-grained global information for superior context comprehension. Among them, fine-grained local information refers to the microscopic and specific information within an image, such as texture and profile characteristics. Coarse-grained global information reflects the macroscopic context of the image, encompassing larger geographical scopes or the overall regional imagery. Global information assists in understanding the positioning and implications of fine-grained targets within a broader environment. Simultaneously utilizing global and local information allows for more effective identification of small targets. Global information offers a comprehensive context,

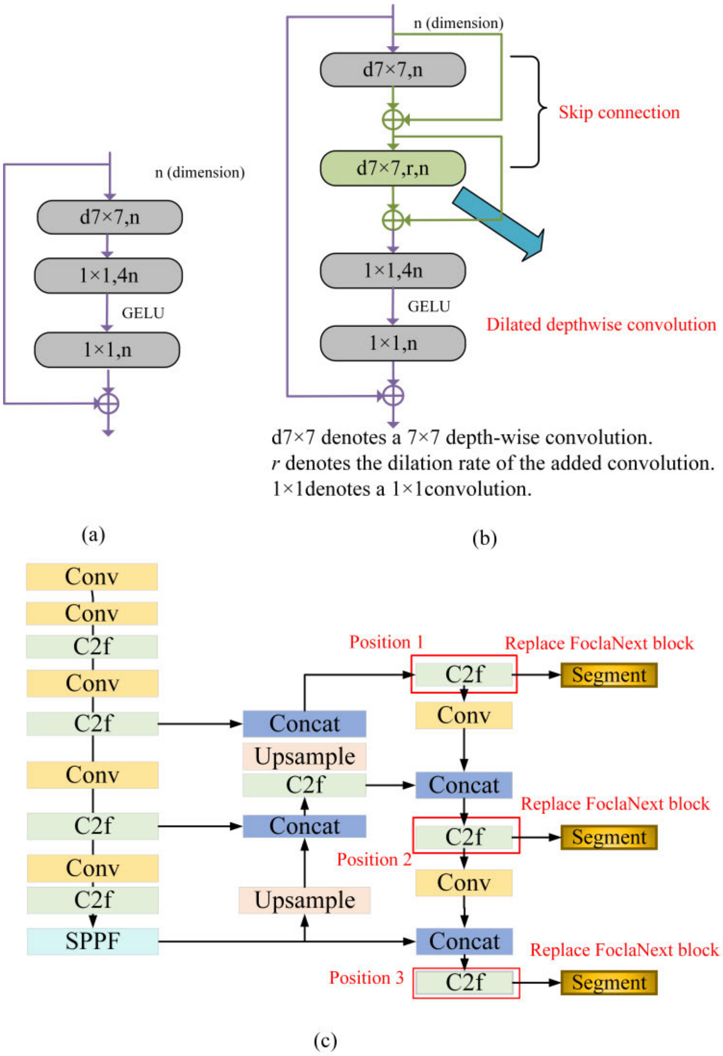


FIGURE 2. The first improvement to the YOLOv8-seg model. (A) ConvNext block, (B) FocalNext block, (C) FocalNext block applied to the YOLOv8-seg model.

allowing the macroscopic understanding of the image, while local information permits precise target identification at the microscopic level.

As depicted in **Fig. 2C**, the C2f module in front of the YOLOV8 segment head is replaced with the FocalNext block module utilizing global and local information fusion.

2) EFFICIENT MULTI-SCALE ATTENTION (EMA)

The Efficient Multi-scale Attention (EMA) module [23] is a novel attention mechanism that builds upon the traditional Coordinate Attention (CA) module [24]. The EMA module refines the representation feature of images by modeling the interaction between channels and spatial dimensions. It does so by efficiently computing the similarity between global and local features, which allows it to capture both long-range and short-range dependencies in the image. The EMA's design allows for a concise and efficient exploration of image features at multiple scales, promoting a better understanding and representation of the underlying structures in the image data. As depicted in **Fig. 3B**, the main procedures of the EMA Attention Mechanism module are as follows:

- Feature Grouping: The EMA module divides the input feature map along the channel dimension, forming multiple sub-features. The approach aids in learning and representing different semantic information. It enhances the model's expressive capacity by thoroughly considering the potential channel correlations and differences.
- Parallel sub-networks: The EMA module employs three parallel paths to extract attention-weight descriptors from the grouped feature maps. Among these paths, two are located at the 1×1 branches, while the third resides at the 3×3 branch. Specifically, two 1D global average pooling operations are utilized to encode the channels within the 1×1 branches (horizontal and vertical directions), as shown in equation (1) and equation (2). The two encoded features are then concatenated and processed through a 1×1 convolution that

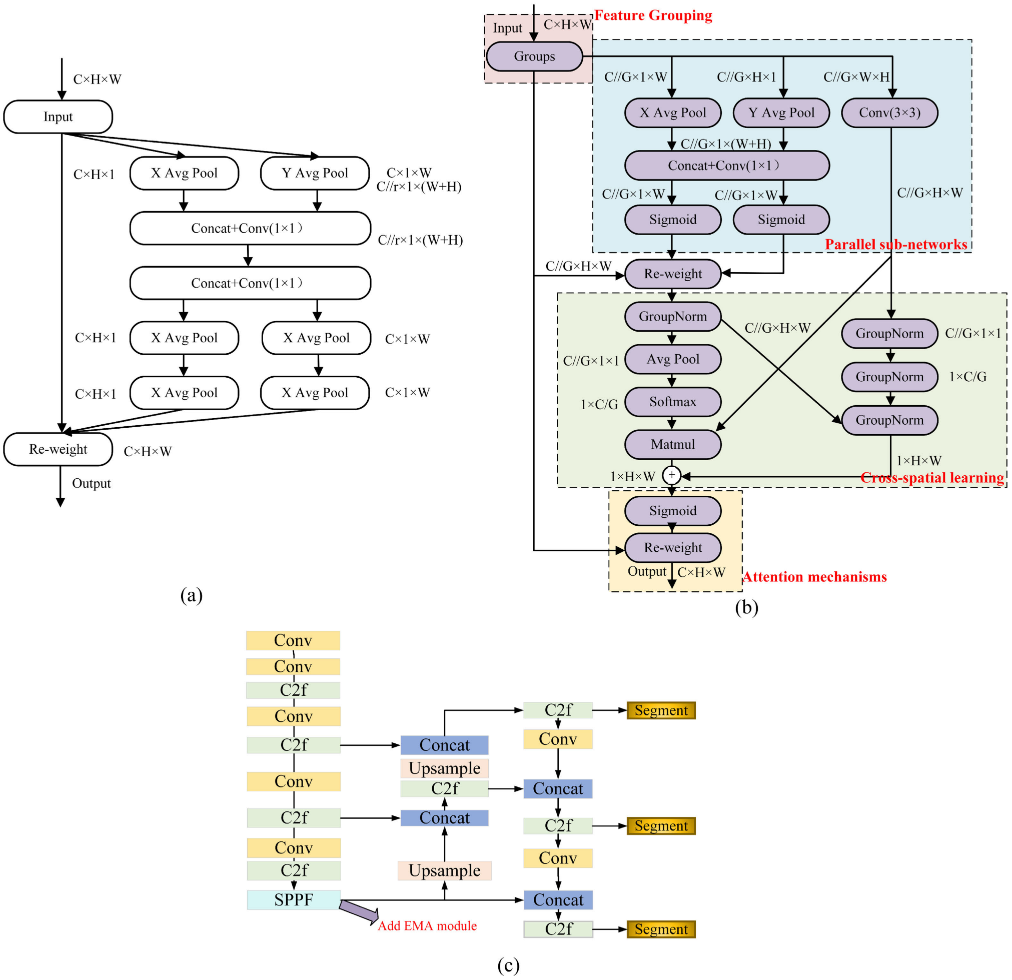


FIGURE 3. The second improvement to the YOLOv8-seg model. (A) CA module, (B) EMA module, (C) EMA module applied to the YOLOv8-seg model.

does not reduce dimensionality. Finally, two-channel attention maps are derived by employing two non-linear Sigmoid functions. For achieving different cross-channel interactive features between the two parallel routes in the 1 × 1 branch, the Parallel sub-networks aggregate the two channel-wise attention maps inside each group via a simple multiplication. As for the 3 × 3 convolution branch, EMA directly applies a 3 × 3 convolution operation on the grouped input feature to enlarge the feature space, then conducts an average pooling on the outcome to generate a second spatial attention map.

$$Z_C^H(H) = \frac{1}{W} \sum_{0 \leq i \leq W} x_c(H, i) \tag{1}$$

$$Z_C^W(H) = \frac{1}{H} \sum_{0 \leq j \leq H} x_c(j, W) \tag{2}$$

where C means the numbers of the input channels, H and W indicate the input features' spatial dimensions, respectively, and x_c indicates the input features at the c-th channel.

- Cross-spatial learning: The EMA module employs a cross-spatial information aggregation method across different spatial dimensions to achieve more enriched feature aggregation. For two spatial attention maps generated by encoding global spatial information within the outputs of the 1 × 1 and 3 × 3 branches, a 2D global adaptive average pooling is first performed on the feature map and reshaped appropriately (as shown in equation (3)), followed by normalization of the result using a softmax function. Finally, the results from both branches undergo element-wise multiplication, producing the first and second spatial attention maps, which retain the full spatial location information.

$$Z_c = \frac{1}{H \times W} \sum_{j}^{H} \sum_{i}^{W} x_c(i, j) \tag{3}$$

where C means the numbers of the input channels, H and W indicate the input features' spatial dimensions, respectively, and x_c indicates the input features at the c-th channel.

- Attention mechanisms: After obtaining the spatial attention maps from each branch, the EMA module employs a non-linear Sigmoid function to generate attention weight values. These values capture pairwise relationships at the

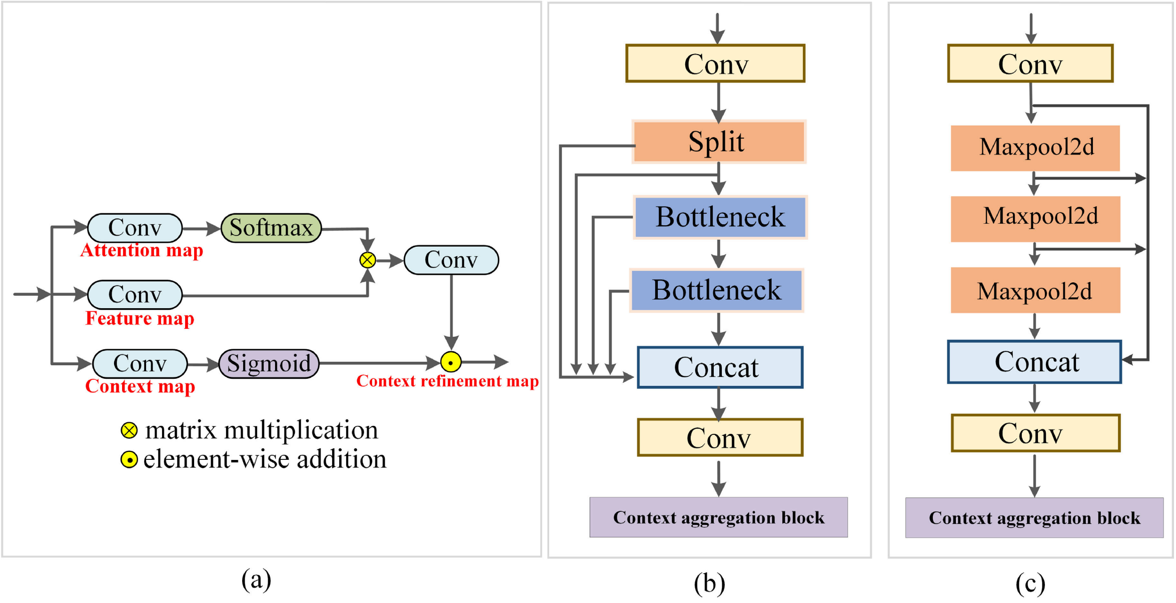


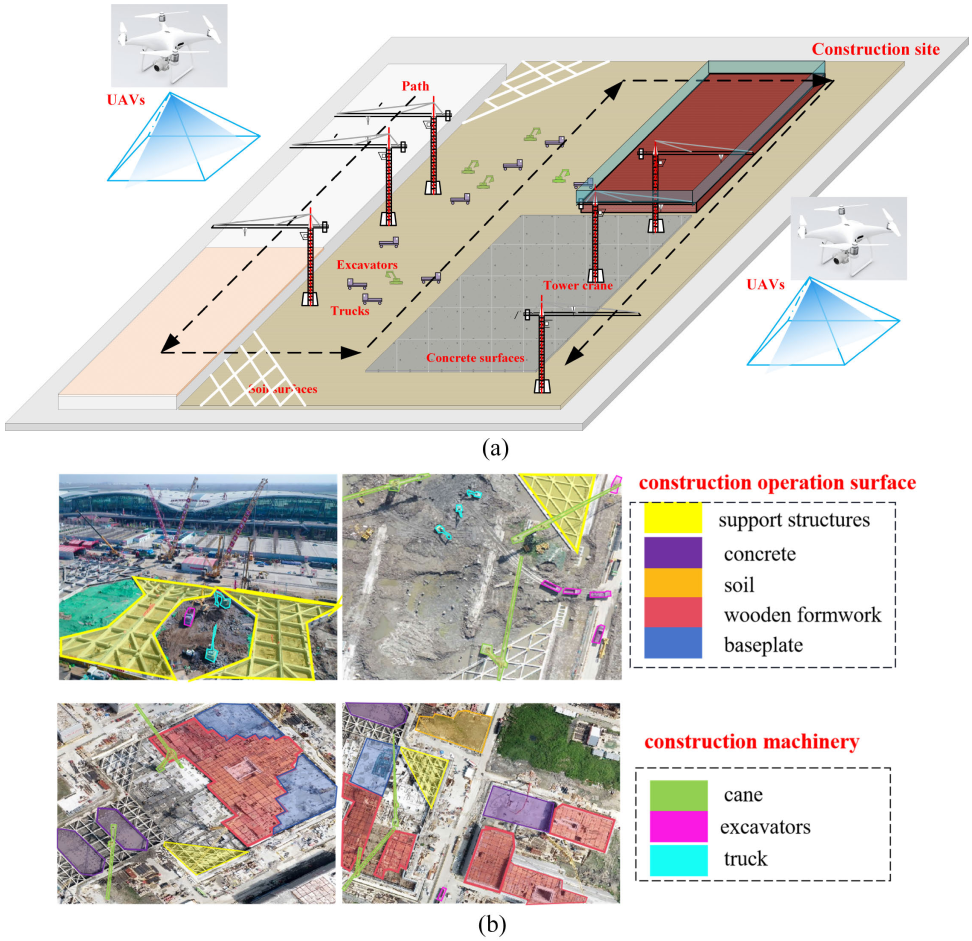
FIGURE 4. The third improvement to the YOLOv8-seg model. (A) Context aggregation block, (B) Improved C2f module, (C) Improved SPPF modu.

pixel level and highlight the pervasive influence of the global context across all pixels. Subsequently, these spatial attention weights are utilized to construct the ultimate feature map.

Computer vision models encounter significant challenges with remote sensing data. The datasets typically include numerous high-resolution images, with small subjects like construction machinery occupying minimal space in the visual field. Detecting these small targets requires robust models capable of discerning their subtle features and separating them from the background and other objects. Moreover, the scale of the same type of construction operating surfaces can appear differently in images due to variations in drone flight altitude and angle. Often these construction operating surfaces display a complex contextual environment requiring a model with advanced capabilities for feature representation and context comprehension. Thus, the feature extraction capability of the model backbone needs improvement. This study adds an EMA attention mechanism module following the model backbone of YOLOv8-seg (**Fig.** 3C). Specifically, the EMA attention mechanism module effectively merges global and local information through global average pooling and dot product operations, allowing the model to capture the global layout and shape information of the construction operating surfaces and understand detailed local features and texture information (such as the details of operating equipment or variations in work progress). Furthermore, the EMA module initially divides the input into multiple sub-features, each with a good distribution of spatial semantic features. The parallel 1 × 1 and 3 × 3 convolutions inside the EMA module enable a more comprehensive capture of the spatial context information of objects. The property aids the model in gaining a deep understanding of complex scenes containing small targets, enhancing the accuracy of small target recognition.

3) CONTEXT AGGREGATION BLOCK

The construction site is a complex and dynamic environment where construction machinery, workers, construction materials, and construction operating surfaces coexist. Given the overlap between construction machinery and operating surfaces, coupled with the intricate nature of the construction environment, distinguishing between these elements presents a significant challenge for accurate object segmentation. Additionally, in remote-sensing image analysis, it has been observed that objects (construction machinery) frequently only take up a small fraction of the total image area, resulting in expansive portions of background information. Traditional convolutional module designs do not adequately distinguish between object and background information, possibly leading to the unnecessary inclusion of excessive non-informative background features. To address the above issue, this paper proposes enhancing the basic modules of YOLOv8 (C2f and SPPF) by incorporating the Context aggregation block [22] to assess each pixel's informativeness for an image, and the improved modules are presented in **Fig.** 4B and **Fig.** 4C. In the Context aggregation block structure, as depicted in **Fig.** 4A, two branches are committed to getting attention and feature maps, respectively. The third branch is devoted to getting the context maps. Then attention map and feature map branches are fused through a matrix multiplication operation. The matrix multiplication operation adaptively weights the input features, thus retaining significant global information while reducing non-pertinent content. Finally, the output result from the matrix multiplication operation is



(a)

(b)

FIGURE 5. The Introduction of the dataset. (A) The process of dataset collection, (B) The dataset annotation categories.

summed element-by-element with the context maps to get the context refinement maps. Specifically, in the Context aggregation block, pixel-wise spatial context is aggregated by:

$$Q_i^j = P_i^j + a_i^j \cdot \sum_{j=1}^{N_i} [\frac{\exp(w_k P_i^j)}{\sum_{m=1}^{N_i} \exp(w_k P_i^m)} \cdot w_v P_i^j] \tag{4}$$

where P_i and Q_i denote the input and output feature maps of level i in the feature pyramid, each containing N_i pixels; j, m ∈ (1) indicate the indices of each pixel; w_k and w_v are linear transform matrices for projecting the feature and context maps (use 1 × 1 convolutions to perform the mapping); a_i is a re-weighting matrix with the same shape as P_i and Q_i to balance the extent of aggregating global spatial context for each pixel, which can also be generated as simple as a linear transform from P_i with softmax normalization, as depicted by equation (5):

$$a_i^j = \frac{\exp(w_a P_i^j)}{\sum_{n=1}^{N_i} \exp(w_a P_i^n)} \tag{5}$$

where j, n ∈ (1) indicate the indices of each pixel; w_a is linear transform matrices for projecting the attention maps (use 1 × 1 convolutions to perform the mapping).

Overall, the Context aggregation block operates under a clear principle: if a pixel's features are informative enough, there is no need to aggregate features from other spatial locations. The approach skillfully balances the integration of critical global contexts with the preservation of unique local features, thereby improving the model's ability to distinguish features in a broad context while maintaining detailed local variation.

III. EXPERIMENTAL INTRODUCTION

A. EXPERIMENTAL DATASET

The study primarily focuses on automatically identifying and monitoring the construction site from a bird's-eye view using an improved YOLOv8-seg algorithm. To accomplish this, we used a drone to collect photographic data from four pit construction project sites and annotate them. The drone utilized for this study was the DJI Phantom 4 Pro. This UAV can achieve top flight speeds of 20 m/s and has a maximum flight duration of around 30 minutes, with a GPS positioning accuracy of vertical ±0.1 m and horizontal ±0.3 m. The onboard camera of the DJI Phantom 4 Pro offers a resolution of 20 megapixels and is equipped with a 1-inch CMOS sensor. Before launching the flight mission, we conducted a comprehensive survey of the construction site to determine the construction operation surface and machinery's position, height, and operational range. We set the drone's flight altitude to 20 meters above the tallest point of the tower crane, allowing for comprehensive coverage and inspection of the entire area. During the actual flight, the drone flew from one end of the construction site, flew in a straight line across to the other end, then turned around and returned along a slightly deviated parallel line. As depicted in **Fig. 5A**, the approach ensures continuous and systematic coverage of the construction area.

Fig. 5B illustrates the captured image samples and their corresponding annotations. Construction sites often do not adhere to a strictly sequential progression of phases because of project scheduling and resource management. Consequently, it is common for several construction stages to be underway simultaneously, leading to scenarios where multiple phases are active simultaneously. As a result, the annotations in the images encompass various construction elements. Finally, the number of different instance categories in the training, testing, and validation sets are summarized in **Table 1**.

TABLE 1. The number of different instance categories in the training, testing, and validation sets.

Category	Training	Validation	Testing
Soil	708	109	228
Baseplate	956	124	276
Wooden formwork	603	83	171
Concrete	1387	196	404
Support structure	572	93	176
Truck	1297	194	388
Cane	664	95	194
Excavator	699	83	174

B. EVALUATION METRIC

Assessing the performance of instance segmentation algorithms needs the utilization of robust error evaluation metrics: Intersection over Union (IoU) refers to the metric that calculates the ratio of the area of intersection to the area of union between the predicted segmentation and ground truth labels; Precision is a statistical metric representing the proportion of positive identifications in a dataset that was indeed correct.; Recall, also known as sensitivity, represents the proportion of actual positives that were identified correctly; The mean Average Precision (mAP) considers precision and recall over varying IoU thresholds, which is particularly well-suited to instance segmentation tasks as it provides a nuanced view of model performance. The formula used to calculate this is provided below:

$$precision = \frac{TP}{(TP + FP)} \tag{6}$$

$$recall = \frac{TP}{(TP + FN)} \tag{7}$$

$$AP = \int_0^1 p(R)dR \tag{8}$$

$$mAP = \frac{\sum_m AP}{m} \tag{9}$$

Frames Per Second (FPS) is a critical performance metric in object detection algorithms, quantifying the number of image frames the algorithm can process per second. FPS is calculated by taking the reciprocal of the time taken to process one frame (in seconds), that is:

$$FPS = \frac{1}{T} \tag{10}$$

C. EXPERIMENTAL CONFIGURATION

The experimental framework for this investigation was constructed on an Ubuntu 18.04 operating system, with Python 3.8.13 as the programming language. The computational libraries used were CUDA-11.4 and cuDNN-8.2.2, paired with PyTorch 1.10.2 for machine learning tasks. The hardware utilized an RTX-3090 GPU equipped with 8GB of memory. The processing unit was an Intel(R) Core(TM) i7-6500M CPU, operating at a clock speed of 3.20GHz.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

YOLOv8-seg offers five differently sized pre-trained models: n, s, m, l, and x. Specifically, YOLOv8n-seg is the smallest model, offering the fastest speed. In contrast, YOLOv8x-seg is the most accurate model, yet it operates at a slower speed. Considering the requirement of edge device deployment for the model parameter, the subsequent research specifically aimed at YOLOv8n-seg and YOLOv8s-seg (as shown in **Table 2**).

TABLE 2. Comparison of different scale YOLOv8-seg models.

Model	Depth	Width	Parameters (M)
YOLOv8n-seg	0.33	0.25	3.26
YOLOv8s-seg	0.33	0.50	11.79
YOLOv8m-seg	0.67	0.75	25.89
YOLOv8l-seg	1.00	1.00	42.90
YOLOv8x-seg	1.00	1.25	67.0

A. ABLATION EXPERIMENTION

In this section, an ablation study was conducted to systematically evaluate the contributions of each enhancement strategy to the overall performance of the YOLOv8-seg

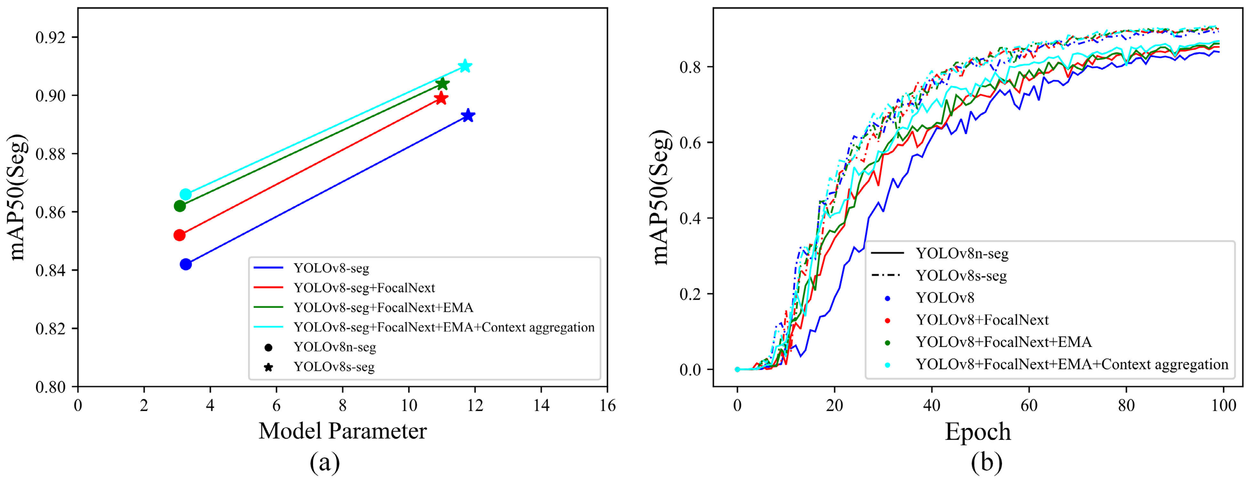


FIGURE 6. Analysis of ablation experiments on segmentation accuracy. (A) Model accuracy - parameter, (B) Validation accuracy changes during training.

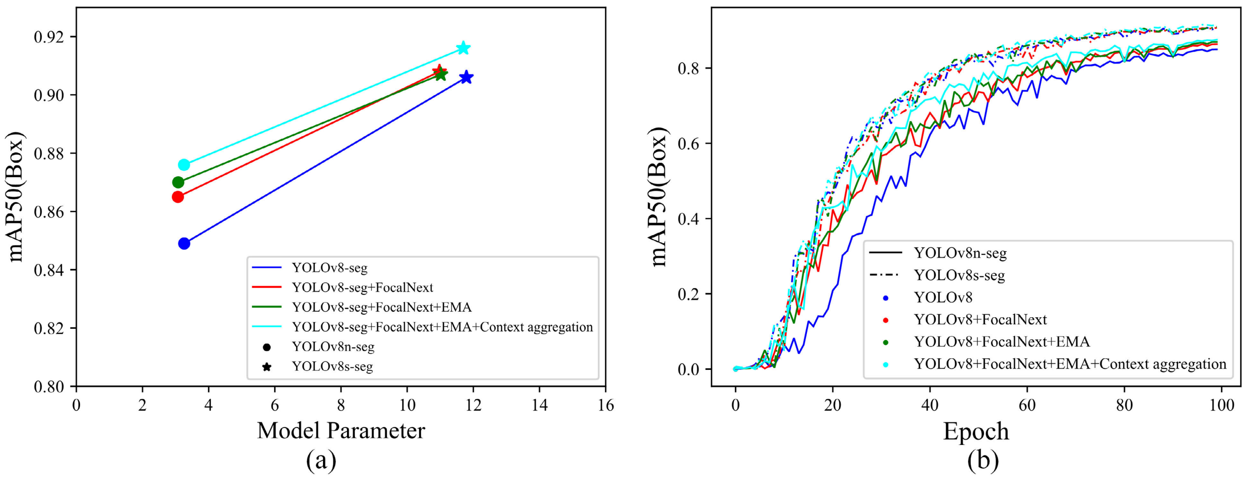


FIGURE 7. Analysis of ablation experiments on detection accuracy. (A) Model accuracy - parameter, (B) Validation accuracy changes during training.

model. To ensure a fair and equitable comparison across all experiments, we maintained a consistent experimental environment: image input size was set at 1200 × 1000 pixels, training epochs were set at 100, SGD was utilized for optimization, and the batch size was fixed at 20, the experimental parameters are listed in **Table 3**. The reasons for choosing the specific values of the hyperparameters are as follows. Image input size: the parameter is obtained based on the average size of the images in our dataset. Training epoch: we observe that the model's performance on the validation set stabilizes after about 100 epochs. The continued increase in the number of epochs does not significantly improve performance. Optimizer: SGD optimizer exhibits greater robustness in the choice of learning rate compared to other optimizers such as Adam or RMSprop. It utilizes only a small batch of samples at a time to gradually adjust and update the model weights, increasing the robustness of the model during the learning process. Batch size: the parameter is chosen based on the memory size of our computer and training time considerations. A batch size of 30 allows us to fully utilize our hardware resources while keeping the training time within acceptable limits.

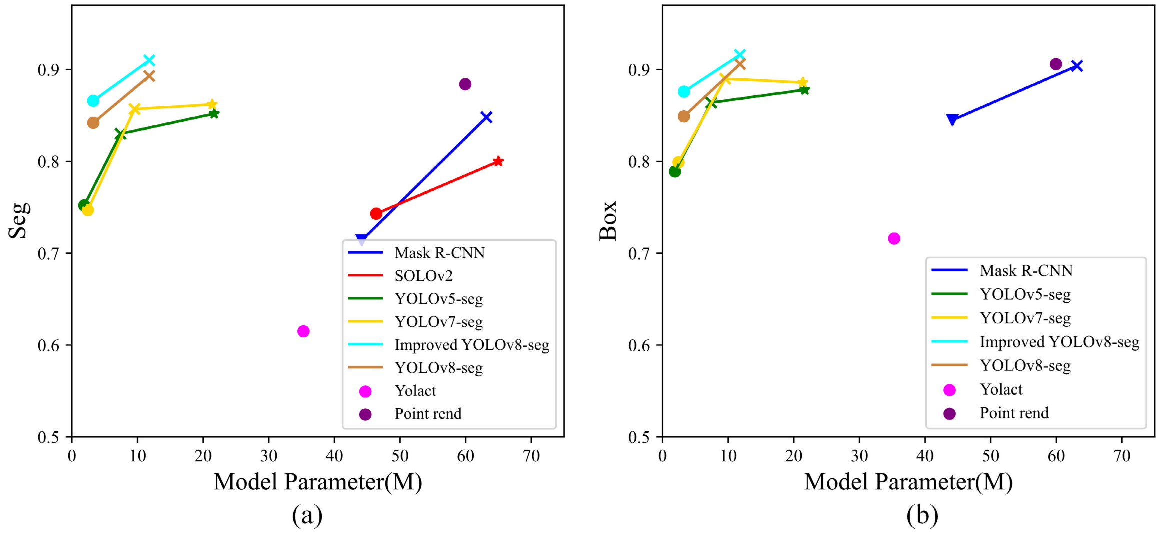
TABLE 3. Experimental parameters.

Experimental parameters	Image input size	Training epoch	Optimizer	Batch size
Value	1200×1000	100	SGD	20

Table 4 summarizes Yolov8-seg models with different improvement strategies: Focal block, EMA, and Context aggravation. Each experiment in the table indicates which improvement strategies were implemented. Experiment 1 served as the baseline (origin YOLOv8), where no improvement strategy was applied, providing a reference for

TABLE 4. Ablation experiments.

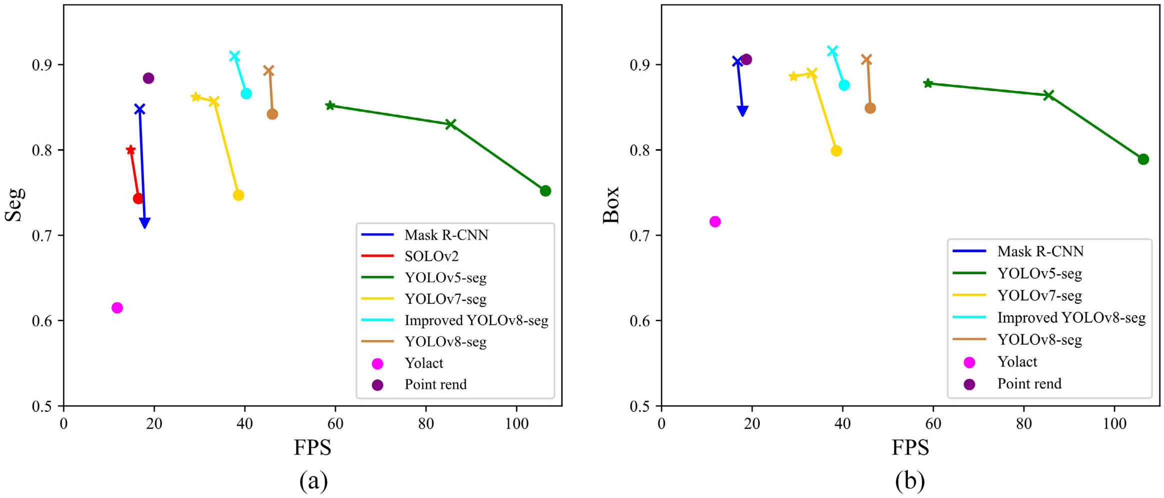
Different YOLOv8-seg algorithm models	FocalNext block	EMA	Context aggregation block
Experiment 1 (baseline)	—	—	—
Experiment 2	√	—	—
Experiment 3	√	√	—
Experiment 4	√	√	√



(a)

(b)

FIGURE 8. Comparison of accuracy and parameter from different instance segmentation models. (A) Seg, (B) Box.



(a)

(b)

FIGURE 9. Comparison of accuracy and FPS from different instance segmentation models. (A) Seg, (B) Box.

evaluating the subsequent experiments. With four independent experiments, we could understand the impact of these strategies on model performance collectively applied.

As depicted in **Fig. 6** and **Fig. 7**, it can be observed that segmentation and detection accuracy show an increasing trend with the gradual increase of the improvement strategies, while the model parameters get reduced compared to the original model. The trend demonstrates the advantages of FocalNext block, EMA, and Context aggregation strategies in enhancing the model performance. It is worth noting that the n and s scale YOLOv8-seg model can observe a similar trend, demonstrating the generalizability and scalability of these strategies. **Table 5** summarizes the effect of the combination of these strategies on the model performance of the n and s

TABLE 5. Ablation experiments.

Ablation experiments	YOLOv8n-seg			YOLOv8s-seg		
	mAP50(Seg)	mAP50(Box)	Parameters	mAP50(Seg)	mAP50(Box)	Parameters
Experiment 1	0.842	0.849	3259624	0.893	0.906	11790472
Experiment 2	0.852	0.865	3069000	0.899	0.908	10976456
Experiment 3	0.862	0.870	3079368	0.904	0.907	11017672
Experiment 4	0.866	0.876	3253924	0.91	0.916	11699572

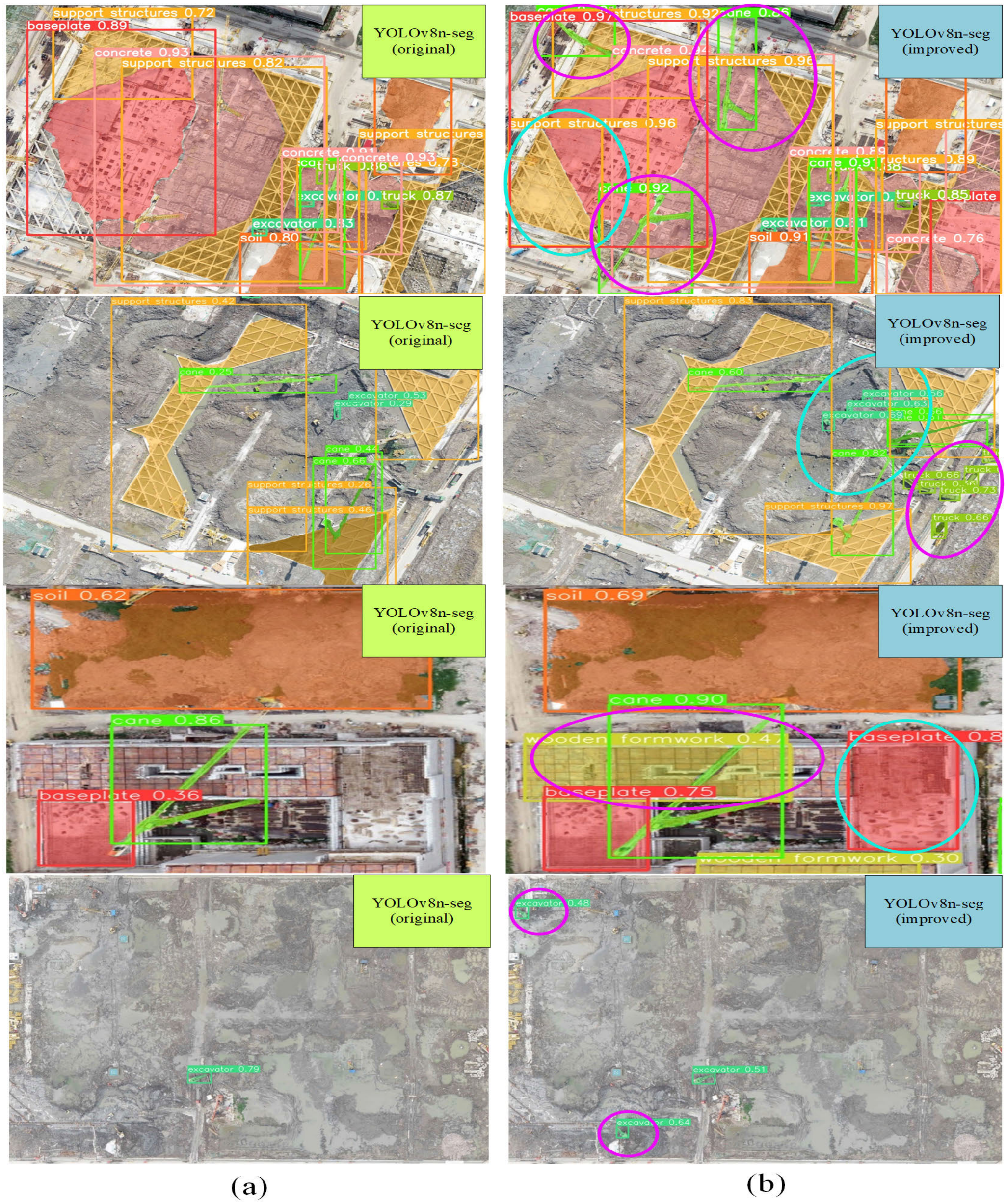


FIGURE 10. The Visualization results of the original YOLOv8n-seg and improved YOLOv8n-seg model in construction site scenes. (A) Original model, (B) Improved mode.

scale models, including the segmentation and detection average precision (mAP50) and the number of model parameters.

Overall, the modifications improved the model's performance and reduced its parameter count, making it an ideal solution

TABLE 6. Comparative experiment.

Object detection algorithm	Backbone or scale	mAP50 (Seg)	mAP50 (Box)	Parameter(M)	T(ms/img)	FPS (img/s)
Mask R-CNN	Resnet50	0.714	0.845	44.17	55.89	17.89
	Resnet101	0.848	0.904	63.16	59.6	16.78
SOLOv2	Resnet50	0.743	-	46.37	60.5	16.53
	Resnet101	0.800	-	65.0	67.5	14.81
Yolact	Resnet50	0.615	0.716	35.29	84.6	11.82
Point rend	Resnet50	0.884	0.906	59.94	53.4	18.73
YOLOv5-seg	n	0.752	0.789	1.89	9.4ms	106.38
	s	0.830	0.864	7.42	11.7ms	85.47
	m	0.852	0.878	21.68	17.0ms	58.82
YOLOv7-seg	n	0.747	0.799	2.43	25.9	38.61
	s	0.857	0.890	9.55	30.1	33.22
	m	0.862	0.886	21.38	34.3	29.15
YOLOv8-seg	n	0.842	0.849	3.26	21.7	46.08
	s	0.893	0.906	11.79	22.1	45.25
Improved YOLOv8-seg (this paper)	n	0.866	0.876	3.25	24.8	40.32
	s	0.910	0.916	11.69	26.5	37.74

for instances of segmentation where high-precision and computationally efficient models are required.

B. COMPARE WITH OTHER MODELS

In this section, the paper establishes a comparative study between the enhanced YOLOv8n-seg model and other prominent instances segmentation models such as YOLOv5-seg, YOLOv7-seg, Mask R-CNN, SOLOv2, Yolact, and Point rend. The comparative experimental results in **Table 6** show that the improved YOLOv8-seg model proposed in this paper has significant advantages over other algorithms regarding instance segmentation mAP50 metrics and model parameters. Regarding the mAP50(Seg) metric, the improved YOLOv8-seg model outperforms all other algorithms in n and s scales. Most prominently, the mAP50 of YOLOv8s-seg reaches 0.910, 2.6% higher than the best comparison algorithm, Point rend. This result indicates that the proposed improved strategy significantly improves the segmentation accuracy of the model. The improved YOLOv8-seg model also shows superiority in the number of model parameters. The model parameters have decreased, significantly smaller than those of other models with the same accuracy (as depicted in **Fig. 8**). In particular, on the s-scale, YOLOv8s-seg has a parameter count of 11.69M, which reduces the model parameters by more than half compared to Point rend (59.94M parameters) and Mask R-CNN (Resnet101) (63.16M parameters), which are similar to mAP50. This means that the improved model significantly reduces the computational complexity and model size while maintaining high accuracy, further improving the model's efficiency. In addition, compared with the previous YOLOv5-seg model and YOLOv7-seg model, the improved YOLOv8-seg model shows a significant improvement in mAP50 with a relatively small increase in the model parameters, proving the effectiveness of the improved strategy. The superiority is attributed to the proposed improvement strategies, such as the FocalNext block, EMA, and Context aggregation block. The improvement strategies simultaneously enhance the model's performance and control the increasement of model complexity to a certain extent, making the model more feasible and applicable in practical applications.

For the speed of model inference, the improved YOLOv8-seg model offers an ideal equilibrium between model performance and speed of inference compared with other models. The speed of the improved model is faster than those of other models with the same accuracy (as depicted in **Fig. 9**). In particular, YOLOv8n-seg achieved an mAP value of 0.866, a competitive accuracy value while maintaining a decent FPS value. Although our improved model slows down in speed compared to the original YOLOv8-seg model and the YOLOv5-seg model, its accuracy is significantly improved. We believe that a good model is about pursuing fast detection speed or high accuracy and finding an optimal balance between the two. Our improved YOLOv8-seg model strives to achieve better results in both aspects.

C. VISUALIZATION RESULT AND ANALYSIS

An image was selected from the test dataset for a comparative analysis of segmentation results. **Fig. 10** shows it after segmentation using the original YOLOv8n-seg and modified YOLOv8n-seg algorithms to intuitively demonstrate the enhanced algorithm's superior performance.

V. CONCLUSION

The main objective of this paper was to improve instance segmentation algorithms for construction site operating surfaces and machinery. The research involving drone photography data collection, data annotation, ablation experiment analysis, and comparative experiment was undertaken.

The research introduced several enhancements to the YOLOv8-seg model, including the FocalNext block, Efficient Multi-scale Attention (EMA), and Context aggregation block strategies. Ablation experiments demonstrated that the gradual Application of these enhancement strategies improves the mAP value while finally reducing the number of parameters of the model compared with the original model. Comparative experiments with other prominent instance segmentation models showed the superiority of the improved YOLOv8-seg model. The performance of the improved model surpassed other algorithms with fewer model parameters and faster inference speed. Overall, this study thoroughly explores the improvement of the instance segmentation algorithms YOLOv8-seg in analyzing construction site data and how the improved instance segmentation algorithms could provide algorithmic solutions for construction site monitoring automation.

VI. LIMITATIONS AND FUTURE WORK

Despite the progress made in our research, some limitations and challenges still need to be further explored.

1. Current instance segmentation models are trained based on a limited number of labeled categories. This limits the model's ability to recognize and process a broader range of building element categories.

2. The dataset in this study focused on light-rich contexts, which may have led to poor model performance for environments with low or varying light conditions.

3. Although the model has shown certain recognition capabilities in the conducted experiments, its robustness in complex and challenging construction scenarios still needs further validation. For example, construction sites with multiple overlapping activities, dense crowds of workers, or large amounts of machinery may pose unique challenges.

Given the limitations described above, future work will focus primarily on the following directions:

1. Research and implement lighting enhancement techniques or preprocessing strategies to enhance image quality in low-light conditions and boost instance segmentation model adaptability to these scenarios.

2. To make the model more widely applicable, we will consider expanding the training dataset to incorporate more construction element categories in the future, thus enhancing the model's recognition capability.

3. In the future, this research will explore adaptive learning mechanisms to make the model more adaptive and robust in the face of highly dynamic, complex, and challenging environments such as construction sites, thereby improving overall efficiency and safety.

ACKNOWLEDGMENT

The authors thank the reviewers who remained anonymous for their constructive criticism and recommendations, which helped improve the manuscript.

DATA AVAILABILITY

The data supporting the conclusions of this article are included within the article. Any queries regarding these data may be directed to the corresponding author.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[2] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9196–9205.

[3] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9796–9805.

[4] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 649–665.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[6] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.

[7] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, and M. Jain, *Ultralytics/YOLOv5: V7. 0-YOLOv5 Sota Real-Time Instance Segmentation*. Honolulu, HI, USA: Zenodo, 2022.

[8] L. Cao, X. Zheng, and L. Fang, "The semantic segmentation of standing tree images based on the YOLO v7 deep learning algorithm," *Electronics*, vol. 12, no. 4, p. 929, Feb. 2023.

[9] W. Yu and M. Nishio, "Multilevel structural components detection and segmentation toward computer vision-based bridge inspection," *Sensors*, vol. 22, no. 9, p. 3502, May 2022.

[10] B. Xiao, H. Xiao, J. Wang, and Y. Chen, "Vision-based method for tracking workers by integrating deep learning instance segmentation in off-site construction," *Autom. Construct.*, vol. 136, Apr. 2022, Art. no. 104148.

[11] K.-S. Kang, Y.-W. Cho, K.-H. Jin, Y.-B. Kim, and H.-G. Ryu, "Application of one-stage instance segmentation with weather conditions in surveillance cameras at construction sites," *Autom. Construct.*, vol. 133, Jan. 2022, Art. no. 104034.

[12] P. Kumar, A. Sharma, and S. R. Kota, "Automatic multiclass instance segmentation of concrete damage using deep learning model," *IEEE Access*, vol. 9, pp. 90330–90345, 2021.

[13] X. Fang, Q. Li, J. Zhu, Z. Chen, D. Zhang, K. Wu, K. Ding, and Q. Li, "Sewer defect instance segmentation, localization, and 3D reconstruction for sewer floating capsule robots," *Autom. Construct.*, vol. 142, Oct. 2022, Art. no. 104494.

[14] S. Siebert and J. Teizer, "Mobile 3D mapping for surveying earthwork projects using an unmanned aerial vehicle (UAV) system," *Autom. Construct.*, vol. 41, pp. 1–14, May 2014.

[15] H. Xie, X. Jing, Z. Sun, Z. Ding, R. Li, H. Li, and Y. Sun, "Tree crown extraction of UAV remote sensing high canopy density stand based on instance segmentation," *Fore. Res*, vol. 35, no. 5, pp. 14–21, 2022.

[16] Y. Song, "Research on vehicle re-identification methods based on UAV aerial images," M.S. thesis, School Control Sci. Eng., Shandong Univ., 2021.

[17] Z. Wang, "Tilted UAV image traffic sign extraction based on Mask R-CNN," M.S. thesis, School Remote Sensing Inf. Eng., Wuhan Univ., Wuhan, China, 2020.

[18] E. L. Stewart, T. Wiesner-Hanks, N. Kaczmar, C. DeChant, H. Wu, H. Lipson, R. J. Nelson, and M. A. Gore, "Quantitative phenotyping of northern leaf blight in UAV images using deep learning," *Remote Sens.*, vol. 11, no. 19, p. 2209, Sep. 2019.

[19] J. Weyler, J. Quakernack, P. Lottes, J. Behley, and C. Stachniss, "Joint plant and leaf instance segmentation on field-scale UAV imagery," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3787–3794, Apr. 2022.

[20] C.-Y. Liu and J.-S. Chou, "Bayesian-optimized deep learning model to segment deterioration patterns underneath bridge decks photographed by unmanned aerial vehicle," *Autom. Construct.*, vol. 146, Feb. 2023, Art. no. 104666.

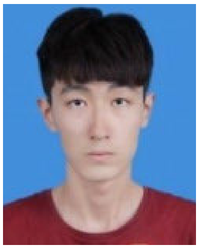
[21] G. Zhang, Z. Li, C. Tang, J. Li, and X. Hu, "CEDNet: A cascade encoder–decoder network for dense prediction," 2023, *arXiv:2302.06052*.

[22] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext V2: Co-designing and scaling convnets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16133–16142.

[23] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[24] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[25] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. W. Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," 2021, *arXiv:2111.11057*.



RUIHAN BAI received the M.S. degree from Hohai University, in 2022. His current research interests include deep learning, computer vision, SLAM, and intelligent construction.

MINGKANG WANG received the B.E. degree in civil engineering from Tongji University, in 2021, where he is currently pursuing the master's degree. His main research interest includes point cloud modeling.

ZHIPING ZHANG received the B.E. degree in civil engineering from Central South University, in 2022. He is currently pursuing the master's degree with Tongji University. His main research interest includes intelligent construction.

JIAHUI LU received the B.E. degree in civil engineering from the Harbin Institute of Technology, in 2022. He is currently pursuing the master's degree with Tongji University. His main research interest includes intelligent construction.

FENG SHEN received the Ph.D. degree in engineering mechanics from Hohai University, in 2014. He is currently an Associate Professor with the School of Civil Engineering, Suzhou University of Science and Technology. He is also a Master Tutor in civil engineering. His research interests include computational mechanics, engineering simulation, and deep learning in engineering structure.