Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 6, June 2026

Research on Weld Defect Recognition Method Based on Mask R-CNN

Yang Enlai, Zhao Zirui

Measurement and Control Technology and Instrument, Jiangsu University, Zhenjiang 212013, China

*Corresponding to: Yang Enlai

Abstract: Automatic detection of weld defects is of great significance to ensuring the quality of industrial products and engineering safety. To address the challenges of low contrast, large defect scale variation, complex background noise, and limited sample data in ultrasonic weld defect images, an improved Mask R-CNN instance segmentation model is proposed. First, the original ResNet+FPN backbone network is replaced with the RSU7 multi-scale feature extraction module to enhance the model's ability to capture details of tiny defects through a nested U-structure and residual connections. Second, the CBAM attention mechanism is connected in series between the backbone network and the region proposal network to suppress background noise and highlight defect regions in both channel and spatial dimensions. Experiments are conducted on an ultrasonic weld defect dataset containing only 105 images. The results show that the improved model achieves a mean average precision (mAP) of 0.7564, which is 4.2% higher than that of the baseline Mask R-CNN (ResNet50+FPN). The coefficient of variation (CV) is 2.26%, indicating better training stability than the comparison models. The maximum AP drop rate under image disturbance is only 9.1%, demonstrating strong robustness. The proposed method realizes high-precision and high-stability defect detection and segmentation in small-sample scenarios, providing an effective technical solution for industrial ultrasonic welding quality inspection.

Keywords: Weld Defect Detection; Mask R-CNN; RSU7; CBAM Attention Mechanism

1 INTRODUCTION

Welding, as a critical joining process in modern industrial manufacturing, is widely used in major equipment sectors such as aerospace, power grids, and petrochemicals. Defects such as porosity, cracks, and lack of fusion are the primary causes of structural failure. Among various non-destructive testing methods, ultrasonic testing (UT) is widely used for weld quality assessment due to its high penetration, high sensitivity, and low cost. However, traditional manual interpretation is heavily influenced by subjective experience, and the resulting inspection efficiency and consistency often fail to meet the automation requirements of smart manufacturing.

With the advancement of deep learning technology, new approaches have emerged for the intelligent identification of welding defects. In the field of welding defect detection, researchers have experimented with various CNN architectures, such as VGG networks [1], ResNet [1], EfficientNet [3], and MobileNet V3 [4]. Researchers including Yu Zeyu introduced an improved LSTM-FCN model, achieving a comprehensive recognition rate exceeding 95.6% for ultrasonic inspection of pipe weld [5]; researchers including Ding Shanzhe employed MobileNet-V2 to identify ultrasonic echo signals from T-joint pipe nodes, achieving an average accuracy of 91% [6]. Xu Xiangqian and colleagues proposed an improved YOLOv5 model for weld surface defect detection, incorporating the self-attention mechanism CoTNet into its Neck layer to effectively reduce redundant information between feature points [7].

Traditional object detection and instance segmentation methods still face the following key challenges when applied to such industrial non-destructive testing tasks: While general-purpose models such as Faster R-CNN and Mask R-CNN demonstrate strong detection capabilities, their backbone networks (e.g., ResNet50) lose high-frequency details of small defects through multiple rounds of downsampling, resulting in insufficient recall for minute pores or microcracks.

To address issues such as low image contrast, wide variations in scale, and susceptibility to noise interference in ultrasonic welding defect images, we propose an improved Mask R-CNN detection model. First, the RSU7 multiscale module is introduced into the feature pyramid network; through a nested U-shaped structure and residual connections, the model's ability to capture details of minute defects is enhanced. Second, the CBAM attention mechanism is incorporated in

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 6, June 2026

series to suppress background noise and highlight defect regions in both the channel and spatial dimensions, respectively. Ultimately, this method is expected to significantly reduce the false negative rate and provide a more accurate instance segmentation solution for industrial ultrasonic welding quality inspection.

2 TECHNICAL FOUNDATIONS OF CBAM AND RSU7

2.1 CONVOLUTIONAL BLOCK ATTENTION MODULE

To enhance the model's focus on defect regions in ultrasonic weld images, the CBAM (Convolutional Block Attention Module) is introduced. This module uses channel attention to adaptively adjust the weights of different feature channels, enabling the network to focus more on defect-related features; it then uses spatial attention to pinpoint the exact location of defects within the image plane, thereby effectively suppressing interference from speckle noise and grain noise in ultrasonic images. As shown in Fig. 1.

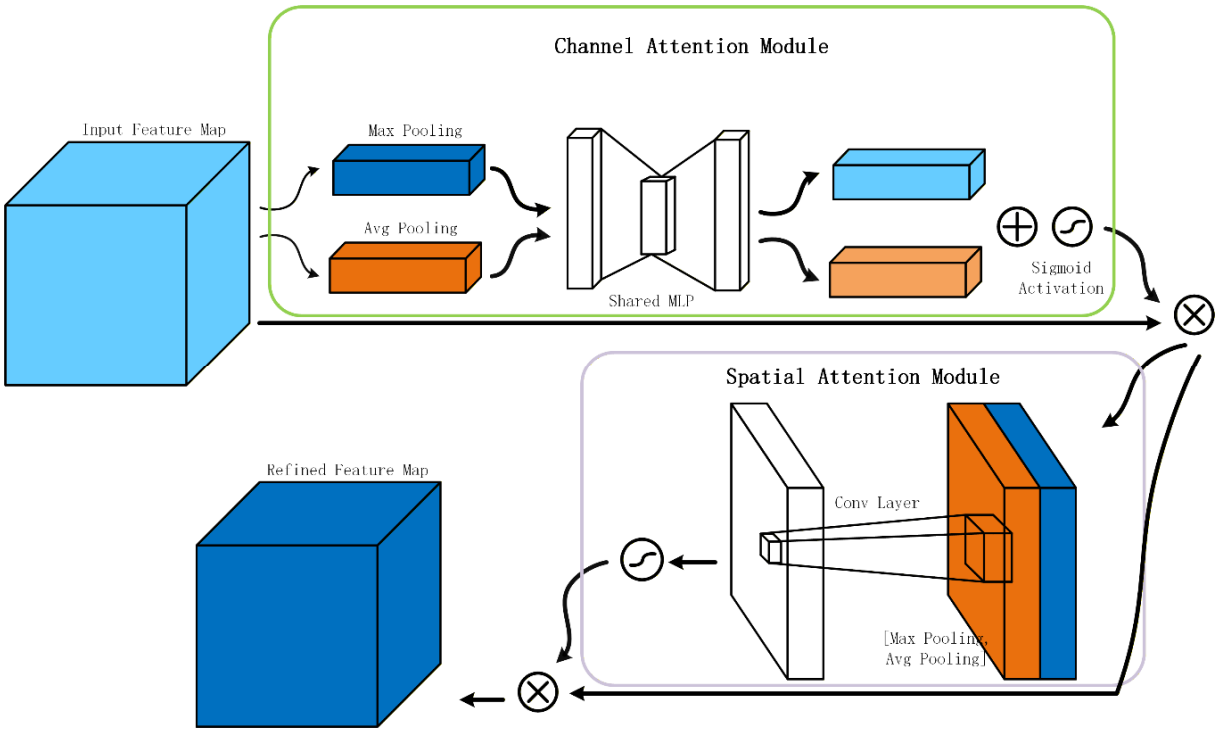

FIG. 1 ARCHITECTURE OF CBAM.

2.1.1 CHANNEL ATTENTION MODULE

The core objective of the channel attention mechanism is to explicitly model the dependencies between channels in the feature map and generate a channel attention map based on these dependencies. As shown in the channel attention module in Fig. 1.

In the specific implementation, this module first reduces the spatial dimension of the input features to one dimension using global average pooling and global max pooling, respectively obtaining two channel descriptors to aggregate global information; subsequently, these two descriptors are fed into a shared multi-layer perceptron for feature transformation to generate the initial channel attention map; Finally, the values of the attention map are normalized to the [0, 1] interval using the Sigmoid function, enabling them to serve as weights for adaptive calibration of the original features.

2.1.2 SPATIAL ATTENTION MODUL

The spatial attention mechanism is designed to explicitly model dependencies between different spatial locations in feature maps, thereby generating a spatial attention map. As shown in the spatial attention module in Fig. 1.

The specific process is as follows: First, the input features undergo max-pooling and average-pooling in the channel dimension to obtain two spatial descriptors. Next, these two descriptors are concatenated along the channel dimension and fused through a convolutional layer to generate an initial spatial attention map. Finally, the values of the attention map are normalized to the range [0, 1] using a sigmoid function, thereby achieving adaptive weighting for different spatial locations.

2.2 RSU7 MULTISCALE FEATURE EXTRACTION

Given the wide variation in defect sizes in ultrasonic welding, the RSU7 module proposed in U²-Net is adopted as the

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 6, June 2026

multi-scale feature extraction unit. This module employs a nested U-shaped architecture that progressively expands the receptive field through seven layers of downsampling, while utilizing residual connections to preserve fine-scale details. This enables the simultaneous capture of defect features across different scales, ranging from microscopic pores to large-scale lack of fusion. Compared to conventional convolutional blocks, RSU7 incurs lower computational overhead because the primary computations are concentrated on low-resolution feature maps [4]. As shown in Fig. 2.

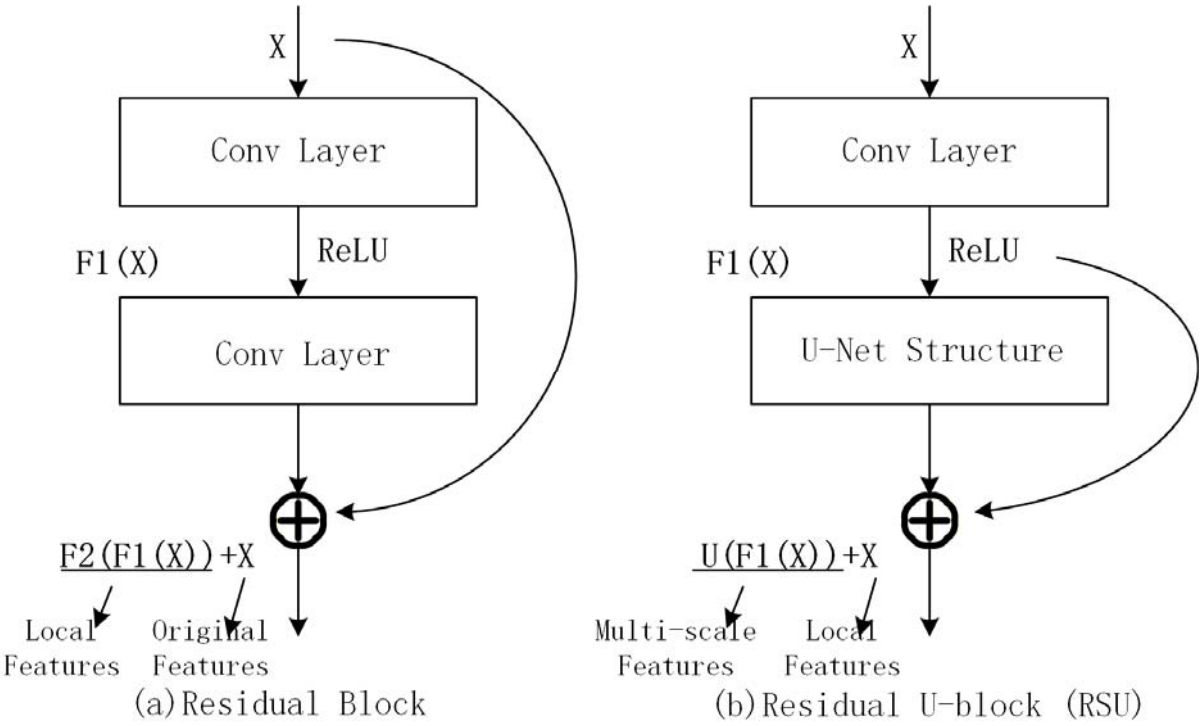

FIG. 2 ARCHITECTURE OF RSU.

The main difference between RSU and the original residual block is that RSU replaces the standard single-stream convolution with a U-Net-like architecture and substitutes the original features with local features transformed through a weighting layer. It is worth noting that RSU has a relatively low computational cost. This is due to its U-shaped structure, and the fact that most computational operations are applied to downsampled feature maps.

3 AN IMPROVED MASK R-CNN DEFECT DETECTION MODEL

3.1 MASK R-CNN OBJECT DETECTION

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 6, June 2026

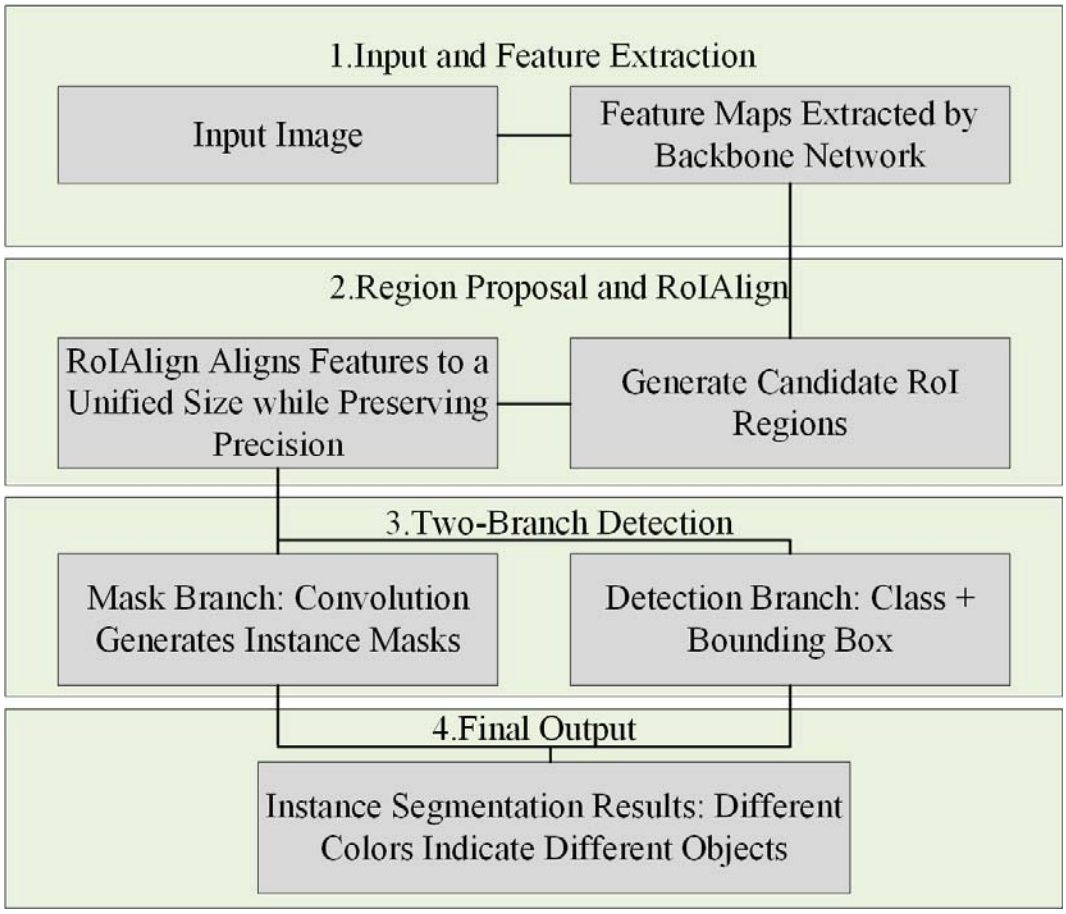

FIG. 3 OVERVIEW OF THE MASK R-CNN ARCHITECTURE.

To address the need for weld defect segmentation, we selected Mask R-CNN as the base framework due to its ability to simultaneously output defect categories, location boundaries, and pixel-level contours [9]. As shown in Fig. 3, this model adds a parallel mask prediction branch to the two-stage detection architecture of Faster R-CNN. In the first stage, the Region Proposal Network (RPN) slides anchor boxes over the feature maps to output candidate defect regions; in the second stage, each candidate region undergoes simultaneous classification, bounding box regression, and mask generatio [9]. To address the feature misalignment issue caused by double quantization in Faster R-CNN's RoI Pooling, Mask R-CNN employs RoI Align, which preserves floating-point coordinates via bilinear interpolation, ensuring that the segmentation boundaries better align with the blurred defect edges in ultrasound images. The backpropagation formula for RoI Align is as follows:

$$\frac{\partial \mathrm{L}}{\partial x_i} = \sum_r \sum_j \delta(i,r,j) \cdot \frac{\partial \mathrm{L}}{\partial y_{rj}} \tag{1}$$

$$\delta(i,r,j) = \begin{cases} (1-\Delta h)(1-\Delta w) & d\big(i, i*(r,j)\big) < 1 \\ 0 & otherwise \end{cases} \tag{2}$$

In the above equation, x_i represents a pixel in the pre-pooling feature map; y_{rj} represents the j pixel in the r candidate region after pooling; and $i*(r,j)$ represents the origin of the pixel value of y_{rj}.

A mask branch is added in the second stage. Mask R-CNN uses a multi-task loss function to jointly train three sub-tasks: classification, bounding box regression, and mask prediction:

$$L = L_{cls} + L_{box} + L_{mask} \tag{3}$$

The definitions of L_{cls} and L_{box} are identical to those in Faster R-CNN. L_{cls} uses a log loss, while L_{box} uses a Smooth L1 loss:

$$smoothL1(x) = \begin{cases} 0.5x^2 & if \ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{4}$$

The mask loss L_{mask} is computed only for RoIs belonging to the positive class, using the average binary cross-entropy loss. For an RoI belonging to class k, the loss is computed using only the kth mask channel; the remaining channels do not contribute:

$$L_{mask} = -\frac{1}{m^2}\sum_{i=1}^{m^2}\left[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)\right] \tag{5}$$

Here, y_i represents the true masked pixel label, and $\hat{y}_i$ represents the predicted probability from the sigmoid output.

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 6, June 2026

This design allows each class to generate masks independently, thereby avoiding competition between classes.

To address the challenges of low contrast, wide variation in defect scale, and complex background noise in ultrasonic welding defect images, we propose two key improvements to the classic Mask R-CNN framework:

1. The RSU7 module is adopted to replace the original ResNet+FPN as the backbone feature extraction network, thereby enhancing the multi-scale receptive field;

2. A CBAM attention module is inserted after the backbone network and before the RPN to suppress background interference and highlight defect regions.

4 Experimental Design and Analysis of Results

4.1 Datasets and Preprocessing

The ultrasound images are 512 × 400 pixels in size, comprising a total of 105 images. These include defect types such as insufficient filling, cracks, porosity, and inclusions; some of these defect types are shown in Fig. 4. When ultrasonic images are applied to machine vision inspection, issues such as low contrast between defects and the background, speckle noise and grain noise interfering with feature extraction, and blurred edges can arise. These problems affect the accuracy of defect localization and dimensional measurement. Furthermore, variations in coupling conditions and probe angles can cause significant image differences, leading to insufficient model generalization ability. Consequently, this impacts the recognition accuracy of convolutional neural network models for defects, posing certain challenges to model identification.

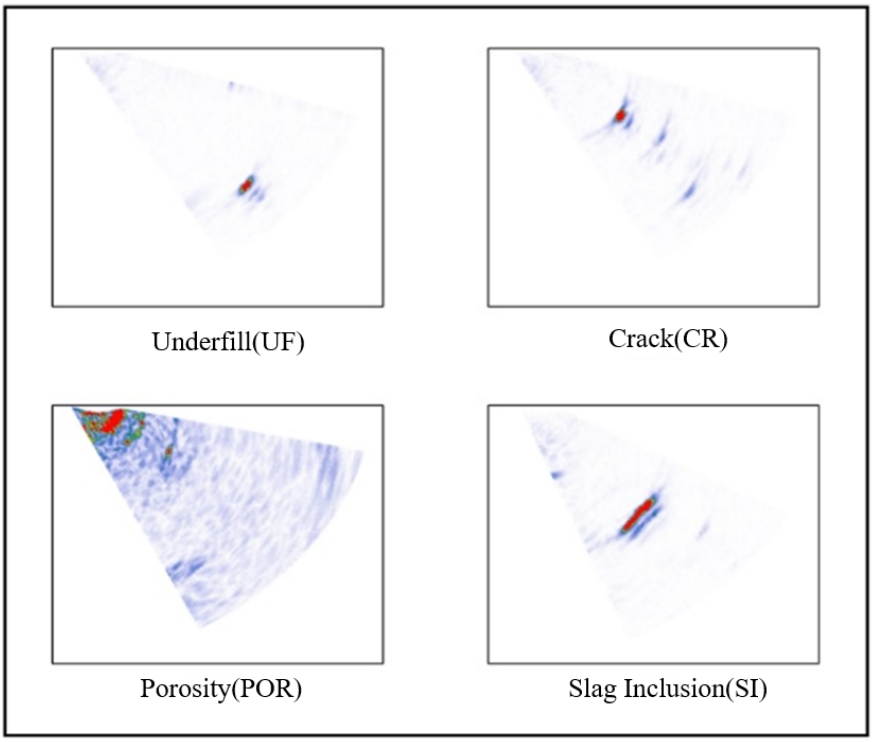

Fig. 4 Ultrasonic Inspection Image.

Model data is annotated using the COCO format, with annotations including defect categories, bounding box coordinates, and masks. The dataset images are split into training, validation, and test sets in a 7:2:1 ratio. During the training phase, data augmentation techniques such as random flipping, rotation, scaling, and noise injection are employed to improve the model's generalization ability.

4.2 Evaluation Criteria

4.2.1 Average Precision (AP)

Following the COCO evaluation framework, Average Precision (AP) is used as the primary evaluation metric. To calculate AP, a series of IoU thresholds (ranging from 0.5 to 0.95 in 0.05 increments) is first set. The area under the precision-recall curve is then calculated for each threshold, and the average of all thresholds is computed. This metric is sensitive to the granularity of predicted boundaries and can assess the model's overall performance under different localization accuracy requirements. Additionally, AP is calculated separately at IoU = 0.50 (denoted as AP_{50}) to evaluate the

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 6, June 2026

model's coarse localization capability—where a prediction is considered correct as long as the overlap area between the predicted and ground-truth bounding boxes exceeds 50%, primarily reflecting the accuracy of determining the presence or absence of defects.

4.2.2 STABILITY

Model stability refers to the consistency of a model's output performance under different training data samples or when the model is run multiple times. Each model is trained five times independently (with parameters randomly initialized each time but using the same training set), with 12 rounds per run. The best mAP value is selected, and the standard deviation of the AP values is calculated to assess the impact of randomness in the training process on model performance. The formula for the coefficient of variation is:

$$CV = \frac{\sigma}{\mu} \times 100 \quad (6)$$

Here, σ represents the standard deviation of the AP across multiple runs, and μ represents the mean of the AP. A smaller CV value indicates that the model's performance is more stable.

4.2.3 ROBUSTNESS

Robustness refers to a model's ability to maintain its detection performance in the face of image disturbances (such as noise, changes in lighting, image degradation, and other interfering factors). In practical industrial ultrasonic inspection scenarios, factors such as changes in transducer coupling conditions, environmental vibrations, and variations in material surface conditions can all cause fluctuations in image quality; therefore, robustness is a key indicator of a model's practical engineering value.

This study employs a method based on the FlagEval evaluation framework to apply natural perturbations to the original images in the test set and assess the degree of performance degradation of each model under these perturbed conditions. The formula for calculating the Robustness Index is as follows:

$$\overline{AP} = \frac{1}{T} \times \sum_{i=1}^{T} \frac{AP_{dist_i}}{AP_{org}} \times 100 \quad (7)$$

Here, APorg represents the model's average accuracy on the original images, AP_{dist_i} represents the average accuracy under the i-th perturbation condition, and T is the total number of perturbation types. This metric reflects the model's ability to maintain its average performance under various types of interference; a higher value indicates better robustness. Additionally, the decline in AP is calculated as follows:

$$\Delta AP = \frac{(AP_{org} - \overline{AP})}{AP_{org}} \times 100 \quad (8)$$

Used to assess the lower bound of the model's performance under the most adverse disturbance conditions.

4.3 RESULT ANALYSIS

To comprehensively evaluate the detection performance, stability, and robustness of each model, five models were compared on the ultrasonic welding defect dataset: the baseline Mask R-CNN (ResNet50+FPN), the baseline with CBAM added, the model with RSU5 replacing the backbone, the model with RSU7 replacing the backbone, and the proposed RSU7+CBAM hybrid model. The primary evaluation metrics include mean average precision (mAP), mAP50, coefficient of variation (CV), and the maximum average precision drop (ΔAP) when input images are perturbed. The specific results are shown in Table 1.

TABLE 1 COMPARISON OF DIFFERENT MODELS.

Model	mAP	mAP50	CV	ΔAP
Mask R-CNN +resnet50	0.7262	0.9500	4.86%	12.9%
Mask R-CNN +resnet50+ CBAM	0.7436	0.9514	3.82%	11.2%
Mask R-CNN+RSU5	0.7098	0.9476	6.65%	15.2%
Mask R-CNN+RSU7	0.7320	0.9546	3.20%	11.8%
Mask R-CNN+RSU7+CBAM	0.7564	0.9490	2.26%	9.1%

In terms of the mAP metric, the improved model achieved a maximum value of 0.7564, representing a 4.2% improvement over the baseline, a 2.4% improvement over RSU7 alone, and a slight improvement over CBAM alone. This indicates that the synergy between RSU7 and CBAM can more effectively extract multiscale features of ultrasonic welding defects and suppress background noise, thereby achieving significant gains in overall localization and classification accuracy. It is worth noting that the model's mAP50 is slightly lower than that of RSU7 alone and the baseline, but it remains at a high level.

The coefficient of variation (CV) reflects the model's sensitivity to random initialization and data sampling perturbations. The CV of the improved model is only 2.26%, significantly lower than the baseline's 4.86% and RSU7's 3.20%, making it the best among all compared models. This indicates that after introducing the CBAM attention mechanism, the model is no longer sensitive to parameter initialization and can stably converge to a similar performance level with each training run, which is beneficial for reliable deployment in actual production.

ΔAP represents the maximum drop in AP when the model is subjected to image perturbations (such as noise, blurring, or brightness changes). A smaller value indicates stronger model robustness. The improved model's ΔAP is 9.1%, far lower than the baseline's 12.9% and RSU5's 15.2%, showing the smallest decline among all comparison models. This indicates that the large receptive field provided by RSU7, combined with

CBAM's adaptive feature re-calibration, enhances the model's tolerance to image quality degradation, enabling it to maintain high detection accuracy even when input images are severely corrupted. While CBAM alone improves robustness (reducing ΔAP from 12.9% to 11.2%), the results are even better when combined with RSU7.

5 CONCLUSION

To address issues such as low image contrast, wide variation in defect sizes, and complex background noise in ultrasonic welding defect images, we propose a method for weld defect detection and segmentation based on an improved Mask R-CNN. This method replaces the original ResNet+FPN backbone network with an RSU7 multi-scale feature extraction module, enhancing the model's ability to capture details of minute defects. Additionally, a CBAM attention mechanism is cascaded between the backbone network and the region proposal network to suppress background interference in both the channel and spatial dimensions, thereby highlighting defect regions. Experimental results show that the improved model achieved an mAP of 0.7564 on a small-sample dataset, representing a 4.2% improvement over the baseline model; the coefficient of variation (CV) was reduced to 2.26%, demonstrating excellent training stability; under image perturbation conditions, the maximum AP drop was only 9.1%, indicating significantly superior robustness compared to the baseline model. It is worth emphasizing that this dataset is small in scale, features low defect contrast, and suffers from severe noise interference, representing a typical low-sampling-number challenge. Yet, the improved model still achieves stable performance gains under these conditions, fully demonstrating the strong feature extraction capabilities and excellent sample efficiency resulting from the synergy between RSU7 and CBAM. This provides a high-precision, highly stable instance segmentation solution for industrial ultrasonic welding quality inspection. Future work will involve further expanding the ultrasonic welding defect dataset by incorporating more samples from real-world operating conditions to validate the model's generalization capabilities. We will also explore lightweight network architectures to improve inference efficiency while maintaining accuracy.

REFERENCES

[1]Wang Guoming, Li Miaomiao. Research on Few-Shot Image Classification Method Based on VGG Network [J]. Computer Knowledge and Technology, 2024, 20(17): 6-10.

[2]Chen Y, Shen H, Zhang G, et al. Identification of weld defects from ultrasonic signals using GASF and an improved DCGAN-ResNet network[J]. Nondestruct Test Eval, 2025, 40(7): 2841-2867.

[3]Yao J, Liu H, Ye J. Enhanced EfficientNet-B0 with dual attention mechanisms for food category classification in X-ray images[J]. J Nondestruct Eval, 2026, 45(2): 55. DOI: 10.1007/s10921-026-01348-4.

[4]Ren Hui, Xia Jing, Lu Jinling, et al. Fault Diagnosis Method for Photovoltaic Modules Based on Infrared Imaging and Improved MobileNet-V3 [J]. Acta Energia-Sinica, 2023, 44(8): 238-245.

[5]Yu Zeyu, Yuan Hongqiang, Wei Xiaolong, et al. Deep Learning-Based Ultrasonic Defect Identification Method for Pipeline Welds [J]. Science and Technology and Engineering, 2022, 22(30): 13288-13292.

[6]Ding Shanze, Ma Xiaochun, Zhang Dezhi, et al. Ultrasonic Identification of Weld Defects Based on Convolutional Neural Network [J]. Acoustics and Vibration, 2024, 13(1): 110-118.

[7]Xu Xiangqian, Li Xing, Zhang Yong'an. Improved YOLOv5 Weld Surface Defect Detection Algorithm [J]. Journal of Xi'an Shiyou University (Natural Science Edition), 2025, 40(4): 98-106.

[8]Xiang L, Xianjin F, Gaoming Y, et al. TransU²-Net: An effective medical image segmentation framework based on transformer and U²-Net[J]. IEEE J Transl Eng Health Med, 2023, 11: 441-450.

[9]Ou Pan, Lu Kui, Zhang Zheng, et al. Target Recognition and Spatial Positioning Based on Mask RCNN [J]. Computer Measurement and Control, 2019, 27(6): 172-176.

[10]Jiang Z, Fu J, Zeng T, et al. Defect R-CNN: A novel high-precision method for CT image defect detection[J]. Appl Sci, 2025, 15(9): 4825.