EGSNet:A Multi-scale Tooth Surface Defect Segmentation Algorithm Combined with Edge Guidance

Jingyi Du, Yifan Bao, Rui Gao, Mengcong Liu, Chenlu Guo, Leqi Li

Abstract—**Accurate segmentation of tooth surface defects in wind turbine gearboxes is essential for ensuring reliable operation and maintenance of wind turbine systems. To solve the problem that the tooth surface defects of wind turbine gearboxes are similar to the background and there are many small-sized defects, this work presents a tooth surface defect segmentation algorithm, EGSNet. First, a lightweight initial structure, DeepStem, is introduced to construct a progressive feature extraction path by stacking multiple 3×3 convolutional layers, which effectively enhances fine-grained representations in shallow layers and improves the perception of tiny and low-contrast defects. Second, a Boundary Perception Module (BPM) is devised to deeply fuse high-level semantic and low-level spatial features from the Feature Pyramid Network (FPN), using attention mechanisms and multi-scale deformable convolutions to adaptively capture complex boundary features, improving the modeling and delineation of defect edges. Finally, an edge-guided loss function based on the Sobel operator is constructed to extract gradient information from multiple directions and impose pixel-level alignment constraints between the predicted mask and the ground-truth boundaries in the loss function, thereby improving the accuracy and clarity of edge segmentation. Experiments are conducted on specialized tooth surface defect data. The results showed that the mAP_{75} and mIoU of the EGS algorithm proposed in this paper are 81.10% and 79.24%, respectively, representing improvements of 3.3% and 1.5% compared to the original network. This validated the effectiveness and practical value of the algorithm in tooth surface defect segmentation tasks.**

Index Terms—**Tooth Surface Defects, DeepStem, BPM, Sobel**

Manuscript received May 7, 2025; revised August 29, 2025.

This work was supported by the Natural Science Special Project (24JK0551) of the Department of Education of Shaanxi Province, China, and the Science and Technology Project (HNKJ24-H86) of China Huaneng Group.

Jingyi Du is a Professor of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: 000248@xust.edu.cn)

Yifan Bao is a postgraduate student in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (Corresponding author, e-mail: pp15691858948@163.com).

Rui Gao is a senior engineer specializing in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: gaorui@xust.edu.cn).

Mengcong Liu is a postgraduate student in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: lmc17191219093@163.com).

Chenlu Guo is a postgraduate student in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: 18991794673@163.com).

Leqi Li is a postgraduate student in Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: menuie@outlook.com).

I. Introduction

Wind energy, a cornerstone of the global energy transition, accounted for 117 GW of newly installed grid-connected capacity worldwide in 2024, representing the highest annual addition on record [1]. As the service life extends, the likelihood of failures in wind turbine systems rises, while expenses for operation and maintenance may account for 30%~35% of the overall wind power cost [2]. The gearbox, as the wind turbine generator chain's central power transmission hub, is the primary cause of unplanned shutdowns and significant maintenance costs [3][4]. The service state of the tooth surface, which acts as the dynamic load-bearing contact during conjugate meshing, play a crucial role in the gearbox's overall operation and stability [5]. As a result, monitoring the condition of tooth surface defects in gearboxes is not only theoretically significant but also invaluable in engineering practice.

In the detection of gearbox tooth surface defects, manual visual inspection is still commonly employed; however, it is inefficient and highly subjective. Therefore, as technology progresses, approaches for detecting tooth surface defects are generally divided into two types, with the first focusing on physical detection techniques, such as acoustic emission, weak magnetic field detection, and eddy current array technologies [6][7][8], the second relies on machine vision for defect detection, and depending on the stage of development, it falls into two groups: conventional image processing techniques and approaches founded on deep learning [9]. Traditional image processing methods are grounded in classical computer vision techniques, performing image feature extraction, processing, and analysis through predefined algorithms. Traditional approaches are largely based on hand-crafted features and manually defined rules, such as edge detection [10], morphological operations [11], threshold segmentation [12], texture analysis [13]. However, traditional image processing methods suffer from limited feature extraction capabilities, poor adaptability, and low accuracy. They are also highly sensitive to image noise, making them difficult to apply in practical scenarios. In recent years, with the rapid advancement of artificial intelligence, deep learning has achieved notable progress in computer vision, showing clear benefits in detection speed, processing efficiency, and adaptability. It has thus emerged as a research hotspot in tooth surface defect analysis. However, most existing deep learning studies focus on defect detection tasks, such as classification or localization, making it difficult to

accurately delineate pixel-level damage boundaries.

The purpose of image segmentation is to classify each pixel of an image, separating it into distinct regions or objects. Unlike image classification, which maps the entire image to a single label, image segmentation provides more fine-grained information and can generally be divided between semantic and instance segmentation [14]. Semantic segmentation uses fine-grained reasoning to anticipate the label of each pixel in the input image [15]. Ashrafi et al. [16] provided effective technical support for quality control in industrial production by combining the semantic segmentation model and target detection model for segmenting small-sized defects in complex backgrounds. Pan et al. [17] used semantic segmentation using an improved U-Net algorithm to detect surface defects, and the network constructs a coding-decoding structure based on the MBConv module, which reduces network parameters without compromising segmentation accuracy, meeting the demand for real-time detection in industrial scenarios. Shi et al. [18] introduced an industrial surface defect detection approach based on a semi-supervised segmentation framework, which addresses issues of insufficient samples and low data utilization in conventional methods, and offers important theoretical and practical implications. Zuo et al. [19] based their work on the semantic segmentation algorithm SegNet and introduced the DenseNet connection method to achieve pixel-level segmentation of surface damage points, which is a representative improvement in semantic segmentation for industrial defect detection. Instance segmentation can be understood as an extension of semantic segmentation, where individual instances of similar objects are independently segmented at the pixel level. Gao et al. [20] proposed an automated method for detecting defects on track surfaces using the Mask R-CNN framework, evaluating the impact of different backbone networks and learning rates on detection performance, and achieving precise segmentation under varied lighting conditions and defect-dense scenarios. Wang et al. [21] used the Mask R-CNN network to detect surface defects on paper disks, and achieved automatic detection and pixel-level segmentation of paper disk defects, confirming the potential of instance segmentation algorithms for industrial product defect detection. Wen et al. [22] proposed an instance segmentation network, YOLACT++, to detect surface defects such as cracks and stains on magnetic tiles. They enhanced detection robustness and stability by incorporating an attention mechanism and a lightweight network, demonstrating the method's effectiveness and practicality. Fu et al. [23] employed the YOLACT++ algorithm to detect bridge cracks. They improved the activation function to mitigate overfitting during training and pre-trained the model on the COCO dataset using transfer learning, which enhanced the detection of small-sample cracks. This approach improves the model's generalization ability on small datasets, resulting in better crack detection performance. Although existing instance segmentation methods have their advantages in different scenarios, most of them still have deficiencies in small target detection, mask edge alignment and robustness in complex backgrounds, in contrast, Mask R-CNN shows higher segmentation accuracy and stronger generalization by its precise candidate region alignment mechanism and independent mask prediction branch, Liu et al. [24] employed the Mask R-CNN network for defect detection on steel surfaces, achieving precise recognition of multi-class defects such as cracks by conducting both object detection and pixel-level segmentation simultaneously, which effectively enhanced detection accuracy in complex backgrounds. Huang et al. [25] introduced an approach for defect detection leveraging Mask R-CNN, which achieved multi-class defect diagnosis by replacing the main network, incorporating an attention module, and employing a path aggregation network. Wang et al. [26] increased detection accuracy by incorporating new fusion routes into the FPN, Mask R-CNN's backbone network, and proposing new evaluation indexes.

Compared with traditional surface defect detection methods, deep learning-based instance segmentation models can autonomously learn discriminative features, adapt to various types and morphologies of tooth surface defects, and exhibit stronger generalization and robustness. The end-to-end learning framework eliminates the need for manual feature engineering, greatly improving the overall performance of defect detection and segmentation. However, due to complex operating environments, tooth surface images of wind turbine gearboxes often suffer from blurring and low contrast caused by uneven lighting, oil contamination, and the small size of defects. These challenges place higher demands on the accuracy and precision of defect segmentation. Existing methods still face issues such as heavy computational cost, redundant network structures, and insufficient capability to capture edge details and small-scale features, making them difficult to apply directly in real-time wind power equipment monitoring.

To solve the problem that there are many small target defects and similar defects and similar backgrounds in the existing tooth surface defect segmentation methods, this paper designs a tooth surface defect segmentation network, EGSNet, built upon an improved Mask R-CNN framework, with the following main contributions:

1) A lightweight starting structure, DeepStem, is proposed to replace the traditional large convolution kernel with multi-layer 3×3 convolutions to construct fine-grained feature expression paths, which effectively improves the network's perception of small-scale targets and strengthens the model's capability in representing the intricate features of tooth surface defects.

2) A BPM edge perception module is proposed, which fuses high and low-level feature information in the FPN, combining the attention mechanism with multi-scale Convolution operations effectively improves the ability to perceive and express target defect boundaries, and strengthens the model's capability to model the edge information.

3) Based on Mask R-CNN, design the edge guidance loss based on the Sobel operator to extract the image gradient information from horizontal, vertical and two diagonal directions to realize the multi-directional edge-aware constraints and improve the segmentation accuracy and contour retention details of small target defects.

II. Related Work

This section presents the architecture and operating principles of the Mask R-CNN instance segmentation network model, along with its loss function and SE attention mechanism.

A. Mask R-CNN

Mask R-CNN [27] build upon Faster R-CNN [28] to perform instance segmentation. Beyond the detection pipeline, it introduces a parallel mask head that predicts a binary mask for each region of interest (RoI). The overall architecture is shown in Fig. 1, the architecture consists of a backbone for feature extraction, a region proposal network (RPN), an RoIAlign module, and three task heads: classification, bounding-box regression, and mask prediction. The image fed into the network is first passed through the backbone to extract feature maps, on which the RPN proposes candidate regions. Next, the RoI Align module ensures that the spatial locations on the feature map are not lost by performing accurate spatial alignment of the candidate regions. The classification branch and the regression branch are used to predict the category label and location regression for each candidate region, respectively. The mask branch performs fine pixel-level segmentation of instances by predicting the corresponding mask image for each RoI. It uses a convolutional network to generate a binarized mask for each RoI and finally outputs the segmentation results for each instance. In this way, the Mask R-CNN can not only perform target detection but also accurately segment each target instance at the pixel level. This structure enables Mask R-CNN to achieve superior performance in multiple-instance segmentation tasks, particularly effective in handling complex scenarios and multiple objects, while maintaining high flexibility and accuracy.

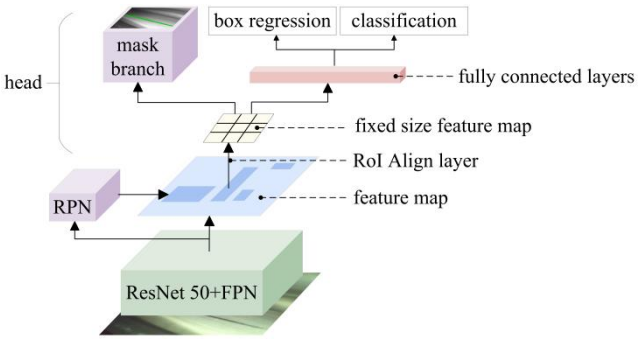


Fig. 1. Mask R-CNN network structure

B. Loss Function

The overall loss function of Mask R-CNN is composed of three components: the classification loss and bounding-box regression loss from the detection branch, together with the mask prediction loss from the segmentation branch, as shown in Equation 1:

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

Where L_{mask} is the mask loss of pixel-level segmentation, L_{box} is the regression loss of the bounding box position, and L_{cls} is the classification loss of the class to which the RoI belongs. L_{cls} employs the standard multiclass cross-entropy loss for quantifying the discrepancy between predicted class probabilities and the true labels, as shown in Equation 2:

$$L_{cls} = -\Sigma_i y_i \log(p_i) \tag{2}$$

Where y_i is the true label of the first class, and p_i denotes the predicted probability. The border regression loss L_{box} is used to minimize the offset between predicted and actual bounding boxes, employing the Smooth L1 loss function to regress the four parameters, including the center coordinates, width, and height, which are defined as shown in Equation 3:

$$L_{box} = \sum_{i\in\{x,y,w,h\}} \mathrm{Smooth}_{L_1}(t_i - t_i^*) \tag{3}$$

Where t_i and t_i^* represent the parameters of the prediction box and the real box, respectively, and the $\mathrm{Smooth}_{L1}(x)$ loss function is defined as shown in Equation 4:

$$\mathrm{Smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

This loss behaves as an L_2 loss with a smoother gradient for small regression errors, and transitions to an L_1 loss for large errors, thus enhancing robustness. Mask R-CNN predicts a pixel-level segmentation mask of the target region on each RoI by a small fully convolutional neural network. This branch uses the binary cross-entropy loss function to perform supervised learning on the mask of each positive sample RoI, and the loss is defined as shown in Equation 5:

$$L_{mask} = \frac{1}{m^2}\sum_{ij}\left[M_{ij}^* \log M_{ij} + \left(1 - M_{ij}^*\right)\log\left(1 - M_{ij}^*\right)\right] \tag{5}$$

Where M_{ij} is the pixel value (in the range [0,1]) that predicts the position in the mask, M_{ij}^* is the true label (0 or 1) of the corresponding position, and the mask size is usually m×m.

C. Squeeze-and-Excitation Attention Module

The SE module, introduced by Hu et al. [29] as a channel-wise attention module, is shown in Fig. 2 and comprises three main components: the Squeeze, Excitation, and Scale operations, where Squeeze and Excitation are two important steps. First, given the input feature map X, an updated feature map is obtained via a standard convolution operation, as defined in Equation 6:

$$F_{tr}: X \to U, X \in R^{W' * H' * C'}, U \in R^{W * H * C} \tag{6}$$

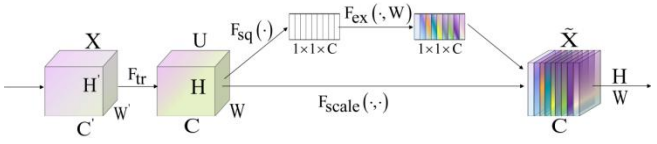


Fig. 2. SE network structure

Global average pooling is applied to the generated feature map to produce a 1×1×C vector, effectively capturing the overall response intensity of each channel. The corresponding equation is given in Equation 7:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{7}$$

Subsequently, the obtained channel descriptor vector is fed into an excitation operation composed of two fully connected layers to model the nonlinear inter-channel dependencies, as defined in Equation 8:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \quad (8)$$

$\sigma(\cdot)$ represents the Sigmoid function. Finally, each channel of the input feature map is multiplied by its corresponding channel weight to achieve dynamic re-calibration of the channel features, as defined in Equation 9:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c \quad (9)$$

III. Algorithm Design

In the task of image segmentation for tooth surface defects in wind turbine gearboxes, the targets are typically small, numerous, and irregularly shaped. Moreover, due to the high similarity between defects and the background, traditional segmentation methods often suffer from false positives and missed detections in such complex scenarios. Although deep learning-based instance segmentation algorithms have improved segmentation accuracy, they still face challenges such as insufficient representation of small target features, loss of edge information, and low feature extraction efficiency. To overcome these limitations, this work introduces a series of improvements built upon the Mask R-CNN framework for instance segmentation. First, for the problem that the 7×7 large convolutional kernel in the first layer of the ResNet-50 backbone network tends to cause the loss of texture information, a starting structure named DeepStem is designed, which adopts three consecutive 3×3 small convolutions instead of the large convolution, and better preserves the edge details of the small defects while keeping the same size of the original features. Secondly, a BPM boundary perception module is proposed for the problem of insufficient edge expression ability of small defective targets. This module integrates high and low-level features in FPN, combines the attention mechanism with Multi-scale convolution operation, and improves the perception and expression ability of small target boundaries. Finally, to further enhance overall segmentation performance, Mask R-CNN's loss function is adjusted to emphasize feature learning and boundary supervision in small target regions, thereby enhancing the model's segmentation performance on small-scale defects. The improvements introduced in this work substantially boost the network's capacity for segmenting minor defects on the tooth surfaces of wind turbines, outperforming the traditional architecture in both accuracy and edge preservation, and demonstrating strong potential for practical application.

A. Design of the BPM Module

In tooth surface defect segmentation tasks, the small size and complex morphology of defect targets often cause traditional feature extraction networks to lose edge clarity and miss fine structural details during representation, thereby limiting segmentation performance in complex backgrounds. In particular, while Feature Pyramid Networks (FPNs) enhance multi-scale feature fusion, their high-level semantic representations offer strong global perception but lack the spatial precision needed to capture fine-grained details. In contrast, while low-level features retain rich edge information, they are easily affected by background noise, making it challenging to precisely delineate defect boundaries. To tackle these challenges, we design a BPM module that improves the network's capability to capture the edges of defect targets, as illustrated in Fig. 3.

The module takes the low-level feature P2 and the high-level feature P5 in the FPN structure as inputs, and adaptively enhances the edge-sensitive channel by introducing the SE attention mechanism, combined with the upsampling operation and feature fusion method, and fully combines the low-level detail structure and the high-level semantic information. On this basis, to improve the network's modeling ability of multi-scale edge structure, multi-scale Dilated Convolution (DC) operations are further introduced. Specifically, three parallel sets of 3×3 dilated convolutions are used, and their dilation rates are set to 1, 2 and 3, respectively, to perceive edge context information at different scales and enhance the response ability to small defect boundaries. Finally, after fusing the above multi-scale features, the edge perception map is generated through a 1×1 convolution and Sigmoid function, which guides the network to focus on the target boundary region of the defect more accurately, to effectively improve the segmentation accuracy. The calculation process is shown below, where $\sigma(\cdot)$ represents the Sigmoid function.

$$F_r = \mathrm{Conv}_{3\times3}^{d=r}\tilde{X} \quad (10)$$

$$F_{DC} = \mathrm{Concat}(F_1, F_2, F_3) \quad (11)$$

$$F_e = \sigma(\mathrm{Conv}_{1\times1}(F_{DC})) \quad (12)$$

Compared to existing mainstream edge modeling methods, the BPM module proposed in this work offers notable advantages in boundary representation. First, the multi-scale dilated convolution structure incorporated in the module enhances the perception of edge context information across various scales, demonstrating strong adaptability in handling small-scale defects. Second, by fusing the middle and high-level semantic features of FPN and the low-level detail features, and combining them with the SE attention mechanism to adaptively adjust the channel information, it is helpful to reduce the background interference while retaining the edge details, to improve the boundary expression ability of the defective target. Furthermore, the edge-aware map generated by the module is integrated into the backbone network as auxiliary information, which guides the network to focus more precisely on defect boundaries within complex backgrounds and contributes to an overall enhancement of segmentation performance.

B. Design of the Backbone Network

The original Mask R-CNN employs ResNet50 as its feature extraction backbone, which uses a large 7×7 convolutional kernel with a stride of 2 in the first layer to achieve rapid downsampling of image features. However, in tooth defect detection tasks, the generally small size of defect targets and the low contrast between defect details and the background often lead this design to lose fine texture information during early feature extraction. This is particularly detrimental to the extraction of edge features for small targets, thereby impairing subsequent segmentation accuracy. To address these issues, the DeepStem initial module is proposed, and its architecture is illustrated in Fig. 4.

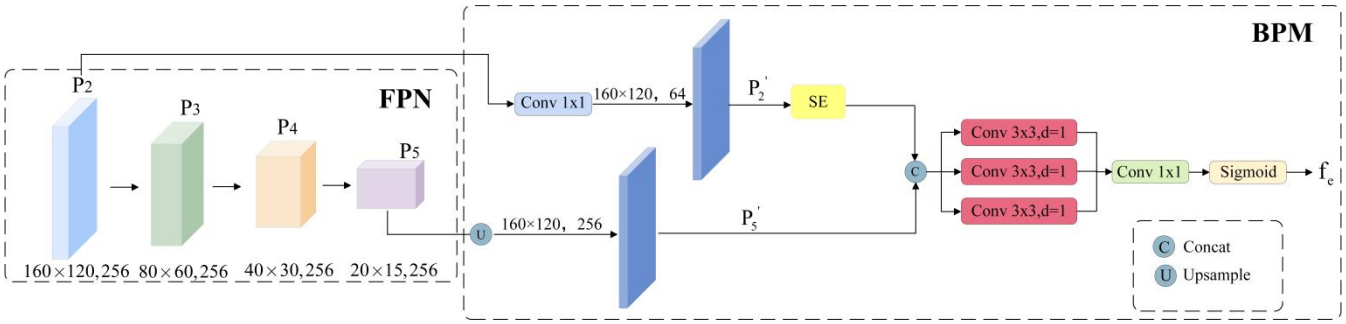


Fig. 3. BPM network structure

The structure replaces the original 7×7 convolution with three consecutive 3×3 convolutional layers. The first two layers focus on fine-grained texture feature extraction, while the third layer employs a stride of 2 to perform downsampling. Stacking multiple small convolutional kernels substantially boosts the network's proficiency in capturing edge and local details. Moreover, this design mitigates the excessive information compression and smoothing typically introduced by large kernels. As a result, it improves the representation of shallow features, particularly for the edges of small defect targets, and provides more discriminative base features for subsequent layers.

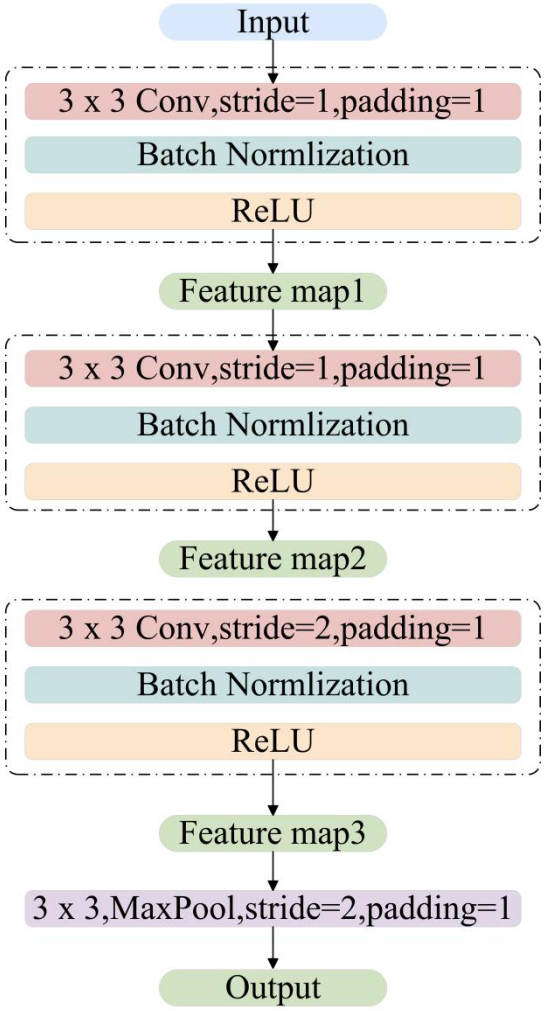


Fig. 4. DeepStem network structure

C. Design of the Loss Function

In gear surface defect detection tasks, traditional loss functions primarily emphasize overall region matching while often overlooking edge features. To address this limitation, this paper introduces an edge-aware loss function, which improves the model's awareness of edge information, thereby enhancing segmentation accuracy and boundary precision. The loss function is built upon the Sobel operator, which captures edge information by computing image gradients in multiple directions. Unlike the standard Sobel method that uses only two convolutional kernels to extract horizontal and vertical gradients, the proposed approach incorporates two additional kernels for diagonal directions, thus enhancing the network's capability to detect edges with greater directional diversity. The improved directional templates and their corresponding convolution kernels are illustrated in Fig. 5.

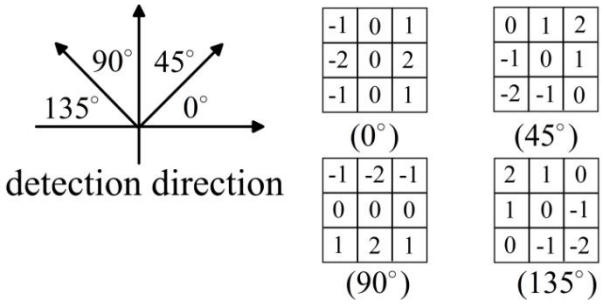


Fig. 5. Improved Sobel operator template diagram

The directional templates shown in the figure indicate the convolution structure of the Sobel operator for the four principal directions. Based on these templates, the edge response at pixel (x,y) is computed by convolving the image I(x,y) with each corresponding directional kernel, as defined in Equation 13:

$$G_\theta(x,y) = I(x,y) * K_\theta, \theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \quad (13)$$

The gradient calculation formula for any pixel is shown in Equation 14:

$$G^* = \sqrt{G_{0^\circ} + G_{45^\circ} + G_{90^\circ} + G_{135^\circ}} \quad (14)$$

The proposed edge loss is shown in Equation 15:

$$L_{edge} = \frac{1}{2}\left(G - G^*\right) \quad (15)$$

Where G is the target edge position of the marker, which is the predicted target edge position, so the improved loss is shown in Equation 16:

$$L_{total} = L_{cls} + L_{box} + L_{mask} + L_{edge} \quad (16)$$

By incorporating a multi-directional edge sensing mechanism, the proposed method enhances the model's capacity to capture target boundary structures, particularly in the segmentation of small-scale and morphologically complex defects. The gradient constraints applied along four directions provide more comprehensive coverage of edge information, thereby supporting improved accuracy and robustness in tooth defect segmentation.

D. Overall Algorithmic Structure

In the defect detection task of wind turbine gearbox tooth surfaces, traditional instance segmentation models exhibit significant limitations in extracting edge features and recognizing small targets. This is primarily due to the typically small size of defect regions, blurred boundaries, and low contrast with the background, which collectively hinder the final segmentation performance. To address these challenges, this paper proposes a synergistically enhanced tooth surface defect segmentation network, EGSNet, built upon the Mask R-CNN framework. The overall model framework designed in this research is shown in Fig. 6.

First, considering that the initial downsampling operation of the original ResNet network is prone to losing small-scale defect texture information, this paper replaces the first layer of ResNet50 with a lightweight DeepStem structure as the front-end feature extraction module. This structure achieves a balance between receptive field expansion and detail retention by replacing large 7×7 convolution kernels with multiple layers of 3×3 convolution kernels. This structure effectively mitigates the loss of edge information for small objects in the early stages of the network and provides more discriminative base features for subsequent advanced semantic modeling. Second, to improve the model's ability to model defective boundaries in complex backgrounds, the BPM boundary perception module is designed. The module integrates shallow detail features and deep semantic features, guides the network to focus on edge-sensitive regions through the attention mechanism, and enhances multi-scale edge perception by combining with dilated convolution, which effectively enhances boundary localization capability and avoids sticking and mis-segmentation issues caused by blurred boundaries. Finally, in terms of loss function design, an edge perception loss function based on the Sobel operator is proposed. By calculating the gradient differences between predicted and ground truth segmentation maps in four symmetric directions (0°, 45°, 90°, 135°), the network is guided to learn structural information at the edges. This loss provides a clear supervisory signal to the boundary region during the optimization process, which enhances the model's responsiveness in the fine-grained boundary region and is an important support to improve the segmentation accuracy and edge clarity. In summary, EGSNet significantly improves the model's capacity to capture small-scale tooth defects and recognize boundaries through synergistic improvements at three levels: feature extraction, edge perception, and loss constraints, which effectively enhances segmentation performance in real-world wind turbine gearbox images.

IV. Experimental Results and Analysis

A. Experimental Dataset

The study of tooth surface defect detection relies on a large number of images of tooth surface defects, but there is no publicly available dataset similar to the images of defects on the gear surface, therefore, this study utilizes a GE industrial endoscope to obtain real defect images of wind turbine gearboxes, and penetrates the endoscope probe deep into the interior of the gearbox from its internal slit to take images of the parts to be inspected, and manually rejects the images of poorer quality, to study the four types of defects, which are. They are rust, unbalance, gluing and crack, as shown in Fig. 7. After filtering, a total of 980 images with a resolution of 640×480 pixels are retained. The distribution of these defect types is summarized in Table I. All images are manually annotated using the Visual Annotation Tool (VIA), and the annotation results are saved in JSON format. Before training, the dataset was randomly partitioned into three subsets: training, validation, and test, following an 8:1:1 distribution.

TABLE I
Various Number of Defects

Defect category	rust	gluing	unbalance	crack
Quantity(Sheet)	278	264	232	206
Percentage of	28.37%	26.94%	23.67%	21.02%

B. Experimental Platform and Parameter Settings

The experiment platform environment as well as the hyperparameter settings in this paper are shown in Table II:

TABLE II
Experimental Platform Environment and Hyperparameter Settings

Designation	Versions/parameters
Operating System	Windows 10
GPU	NVIDIA GeForce RTX4060Ti
VRAM	24G
framework	TensorFlow and Keras
CUDA version	11.6
Python	3.6.0
Epoch	300

C. Evaluation indicators

The proposed model is assessed using standard metrics commonly employed in instance segmentation tasks, including mean Average Precision (mAP), mean Intersection over Union (mIoU), total number of trainable parameters (Params), and model inference speed (FPS). mAP reflects the model's ability to comprehensively account for both precision and recall in instance segmentation tasks, and its calculation is provided in Equation 17:

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \quad (17)$$

Where N represents the total number of categories, and AP denotes the average precision of a specific category at different recall rates, equivalent to the area under the PR-curve with accuracy as the x-axis and recall rate as the y-axis. mAP_{50}, mAP_{75}, and $mAP_{50\text{-}95}$, respectively, indicate

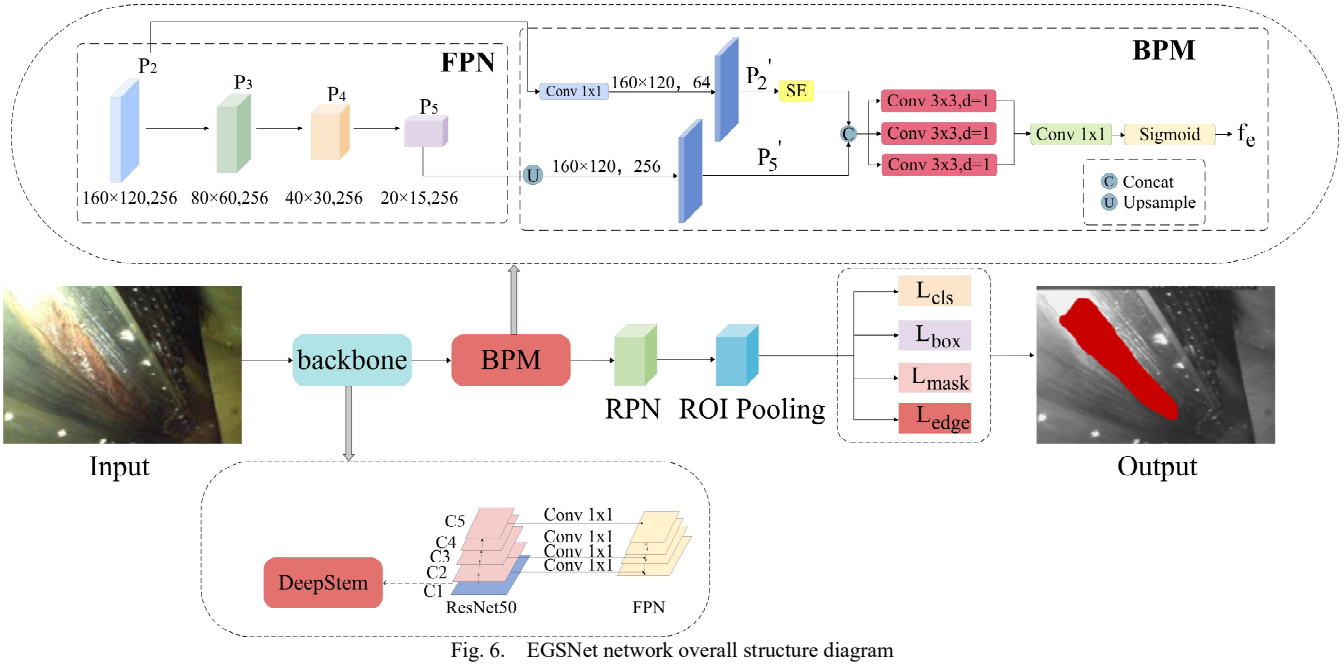


Fig. 6. EGSNet network overall structure diagram

the average precision when the threshold IoU is set to 0.5, 0.75, and 0.5-0.95. mIoU quantifies the overlap between model-predicted segmentation regions and the corresponding ground truth annotations. The calculation process is as shown in Equation 18:

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{IoU}_i \tag{18}$$

Where N represents the number of categories, and IoU represents the intersection-union ratio of a certain category, which represents the intersection-over-union between the predicted and ground truth masks of a single category.

D. Analysis of Experiment Results

1) *Feasibility experiment results and analysis:* Feasibility experiments were performed on the tooth surface dataset to assess the proposed algorithm, and a detailed analysis of its effectiveness was conducted. The evaluation metrics considered include mean average precision mean mAP, instance segmentation metrics mean intersection and merger ratio, mIoU, number of parameters, and detection speed.

Feasibility testing was conducted on the six model groups listed in the table, and the corresponding variations in the primary evaluation metrics are summarized in Table III.

TABLE III
COMPARISON OF FEASIBILITY EXPERIMENT RESULTS

DeepStem	BPM	L_{edge}	mIoU(%)	mAP_{75}(%)
---	---	---	78.06	78.49
√	---	---	78.31	78.51
---	√	---	78.63	78.61
---	---	√	78.15	78.65
√	√	---	78.48	79.19
√	√	√	79.24	81.10

The experimental results demonstrate that integrating the DeepStem and BPM modules into the original network backbone significantly enhances segmentation performance. Specifically, relative to the baseline Mask R-CNN, mIoU and mAP_{75} increased by 0.54% and 0.90%, respectively. Furthermore, the incorporation of an edge loss function further boosts detection accuracy, yielding improvements of 1.5% in mIoU and 3.3% in mAP_{75}. These findings provide comprehensive evidence of the efficacy of the algorithm developed in this study for segmenting gear-tooth defects. As shown in Fig. 8, our algorithm markedly enhances the contrast at defect edges and strengthens edge continuity, with particularly notable improvement for crack defects. This indicates that the model developed in this study is capable of focusing more effectively on defect-critical regions and accurately capturing edge details through enhanced boundary perception. Fig. 9 shows the comparison of the loss curves of the original network and the proposed algorithm during the training process, which shows that the convergence speed of the proposed algorithm is faster and the convergence effect is better.

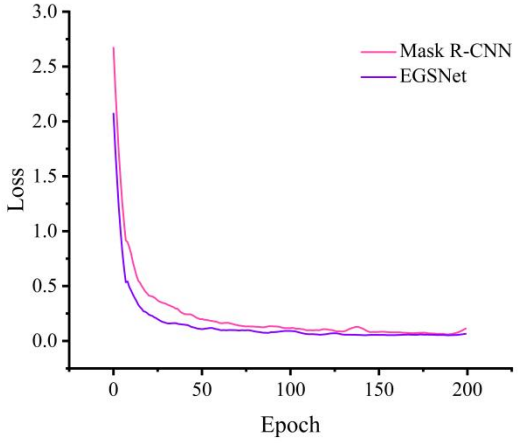
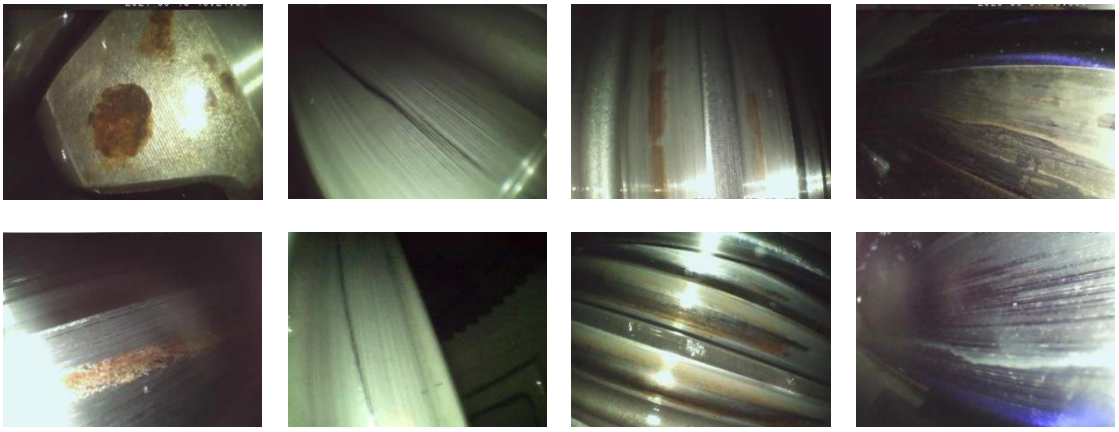
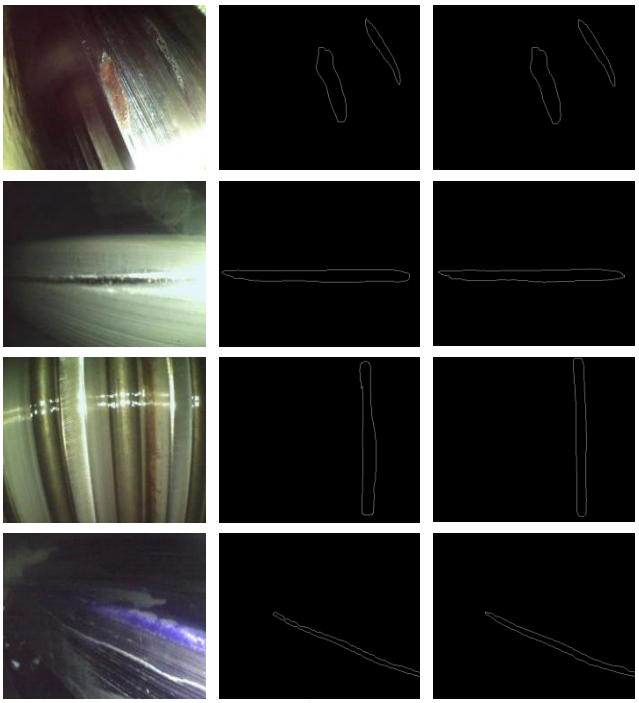


Fig. 9. Training loss comparison between the baseline network and the proposed method in this study



(a) (b) (c) (d)

Fig. 7. Tooth surface diagram:(a): rust;(b): unbalance;(c): gluing;(d): crack



(a) (b) (c)

Fig. 8. Comparison chart of network visualization:(a): Image; (b): Mask R-CNN;(c): EGSNet

2) *Comparative experimental results and analysis*: To comprehensively assess the effectiveness of EGSNet, this study was compared with multiple instance segmentation models, and the evaluation was based on a unified test set, including four types of images: rust, gluing, unbalance, and crack. Table IV shows the experimental results, together with the detailed performance metrics for different instance segmentation models. The structure of the visualization graph is shown in Fig. 10, illustrating how various models perform in segmenting the four defect categories.

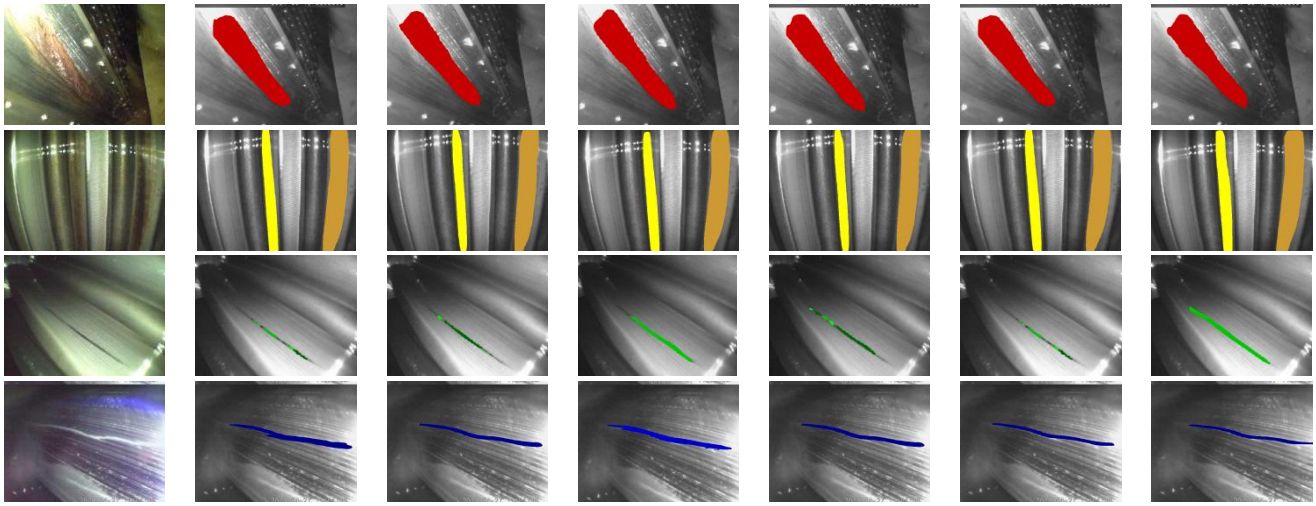
As shown in Table IV, compared with existing instance segmentation algorithms, our algorithm has obvious advantages in mAP_{75} and mIoU metrics while ensuring parameter quantity and inference speed. In addition, the visualization in Fig. 10 indicates that our algorithm performs better in edge detection and overall segmentation.

TABLE IV
COMPARISON OF EXPERIMENT RESULTS

Network	mIoU(%)	mAP_{75}(%)	Params/M	FPS
Mask R-CNN	78.06	78.49	56.7	10.1
YOLACT[30]	78.56	79.32	56.4	10.2
BlendMask[31]	78.18	78.52	57.1	11.4
SOLOv1[32]	74.06	75.26	41.8	8.7
SOLOv2[33]	74.12	76.18	43.9	9.3
EGSNet(our)	79.24	81.10	62.8	12.7

V. CONCLUSION

This paper presents EGSNet, a novel instance segmentation network for wind turbine tooth surface defect detection, built upon the Mask R-CNN framework. The model is designed to address the challenge of identifying small and medium-sized defects that often blend into the background. To enhance the network's sensitivity to such subtle defects, a lightweight DeepStem structure is introduced at the front end. This module replaces the traditional large-kernel convolution with a stack of 3×3 convolutions, enriching the feature extraction pathway and improving the representation of fine-scale defect features. To further improve the ability of the model to capture defect boundary information, a Boundary Perception Module BPM is introduced. This module integrates high-level and low-level features within the FPN framework, leveraging attention mechanisms and multi-scale dilated convolutions to enable fine-grained edge perception. Additionally, to strengthen boundary supervision, an edge-guided loss based on the Sobel operator is incorporated. By extracting gradient information in horizontal, vertical, and diagonal directions, this loss function imposes multi-directional edge constraints, effectively improving contour preservation and segmentation accuracy for small-scale defects. Results indicate superior performance of the proposed method compared with the baseline network in both segmentation accuracy and edge detail preservation, showing strong generalization and practical applicability. The current model is trained and validated on wind turbine gearbox tooth surface images acquired via industrial endoscopes. However, challenges such as limited viewing angles, drastic lighting variations, and sensor noise introduce constraints on its performance under complex real-world conditions. In particular, for crack defects affected by specular reflections



(a) (b) (c) (d) (e) (f) (g)

Fig. 10. Comparison chart of network visualization effect:(a): Image;(b): Mask R-CNN;(c): YOLACT;(d): BlendMask; (e): SOLOv1;(f): SOLOv2;(g): EGSNet

or blurred boundaries, the model's edge representation and segmentation stability remain areas for further improvement. Future research will further enlarge the tooth surface image dataset to cover various operating conditions and diverse scenarios, aiming to improve the model's capability to operate effectively across different settings, including offshore and inland wind farms. Meanwhile, varying lubrication states, including clean surfaces and oil contamination, also affect image features, requiring the model to demonstrate stronger robustness against such differences. In addition, variability in imaging devices, along with differences in shooting angles, lighting conditions, and operational procedures, leads to changes in image style and quality, which consequently affect model performance. Therefore, improving the model's adaptability and generalization to such complex real-world factors is a crucial task before engineering deployment. Future efforts will integrate image preprocessing, defect enhancement, domain adaptation, and transfer learning techniques to improve model stability and accuracy in complex environments, providing more reliable technical support for the health assessment and intelligent operation and maintenance of wind power gearboxes.

REFERENCES

[1] Global Wind Energy Council, "GWEC Releases Global Wind Report 2025: Global Wind Power Installations Hit Record High in 2024," East Wind Power Network, Apr. 23, 2025. [Online]. Available: http://www.eastwp.net/news/show.php?itemid=78954.

[2] H. Luo, L. Wang, W. Sun, and C. Lu, "Intelligent Monitoring and Maintenance of Wind Turbine Blades Driven by Digital Twin Technology," *Proceedings of the 2023 3rd International Conference on New Energy and Power Engineering*, Huzhou, China, Nov. 24–26, 2023, pp. 626-630.

[3] D. Wang, C. Cao, N Chen, W Pan, H. Li, and X. Wang, "A Correlation-Graph-cnn Method for Fault Diagnosis of Wind Turbine Based on State Tracking and Data Driving Model," *Sustainable Energy Technologies and Assessments*, vol. 56, p. 102995, 2023.

[4] T. Wang, Q. Han, F. Chu, and Z. Feng, "Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review," *Mechanical systems and signal processing*, vol. 126, pp. 662-685, 2019.

[5] Z. He, Y. Yang, and D. Liang, "A Multi-concurrent Fault Diagnosis Scheme for the Parallel Shaft Gearbox Based on Resnet Neural Network and Image Recognition Approach," *Proceedings of the China Automation Congress*, Beijing, China, Oct. 22–24, 2021, pp. 6123-6127.

[6] G. Ciaburro and G. Iannace, "Machine-Learning-Based Methods for Acoustic Emission Testing: A Review," *Applied Sciences-Basel*, vol. 12, no. 20, p. 10476, 2022.

[7] S. Zhang, S. Lu, and Xu Dong, "Stress and Corrosion Defect Identification in Weak Magnetic Leakage Signals Using Multi-Graph Splitting and Fusion Graph Convolution Networks," *Machines*, vol. 11, no. 1, p. 70, 2023.

[8] T. Meng, Y. Tao, Z. Chen, J. R. Salas Avila, Q. Ran, Y. Shao, R. Huang, Y. Xie, Q. Zhao, Z. Zhang, H. Yin, A. J. Peyton, and W. Yin, "Depth Evaluation for Metal Surface Defects by Eddy Current Testing Using Deep Residual Convolutional Neural Networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, p. 2515413. 2021.

[9] Z. Huang, C. Zhang, L.e Ge, Z. Chen, K. Lu, and C. Wu, "Joining Spatial Deformable Convolution and a Dense Feature Pyramid for Surface Defect Detection," *IEEE Ttransactions on Instrumentation and Measurment*, vol. 73, p. 5012614, 2024.

[10] E. Manik, "Relationship between Segment Edges and Thresholds on Segmentation Generated by Minimum Spanning Trees," *Engineering Letters*,vol. 28, no 3, pp. 769-802, 2020.

[11] B. Li, S. Cao, C. Xu, and S. Huang, "Surface Defect Detection Algorithm for Printing Roller Based on Global Contrast and Edge Gradient," *Proceedings of the SPIE*, vol. 11913, *Sixth International Workshop on Pattern Recognition*, Beijing, China, June 25–27, 2021.

[12] F. Nie, and P. Zhang, "Threshold Selection with Relative J-Divergence for Image Segmentation," *IAENG International Journal of Applied Mathematics*, vol. 53, no.3, pp. 899-906, 2023.

[13] X. Wang, S. Wu, and Y. Liu, "Detecting Wood Surface Defects with Fusion Algorithm of Visual Saliency and Local Threshold Segmentation," *Proceedings of the Ninth International Conference on Graphic and Image Processing*, Qingdao, China, Oct. 14–16,2017, 2018.

[14] R. Ruksana, J. R. Jim, M. J. Hossain, A. Das, M. M. Kabir, and M. F. Mridha, "From Image Classification to Segmentation: A Comprehensive Empirical Review," *Iran Journal of Computer Science*, vol. 8, pp. 271-301, 2025.

[15] A. M. Hafiz, G. M. Bhat, "A Survey on Instance Segmentation: State of the Art," *International Journal of Multimedia Information Retrieval*, vol. 9, pp. 171-189, 2020.

[16] S. Ashrafi Vayghan, S. Teymouri, S. Etaati, J. Khoramdel, Y. Borhani, and E. Najafi, "Steel Surface Defect Detection and Segmentation Using Deep Neural Networks," *Results in Engineering*, vol. 25, p.103972, 2025

[17] J. Pan, D. Zeng, Q. Tan, Z. Wu, and Zhigang Ren, "EU-Net: A Novel Semantic Segmentation Architecture for Surface Defect Detection of Mobile Phone Screens," *IET Image Processing*, vol. 16, no. 9, pp. 2568-2576, 2022.

[18] C. Shi, K. Wang, G. Zhang, Z. Li, and C. Zhu, "Efficient and Accurate Semi-Supervised Semantic Segmentation for Industrial Surface Defects," *Scientific reports*, vol. 14, no. 1, p. 21874, 2024.

[19] Z. Zuo, X. Wang, Y. Wang, B. Wang, S. Wei, and S. Dai, "Armor Damage Point Segmentation Based on Improved SegNet," *Engineering Letters*, vol. 33, no. 6, pp. 1983-1991, 2025.

[20] F. Guo, Y. Qian, D. Rizos, Z. Suo, and X. Chen, "Automatic Rail Surface Defects Inspection Based on Mask R-CNN," *Transportation research record*, vol. 2675, no.12, pp.655-668, 2021.

[21] X. Wang, Y. Gao, J. Dong, X. Qin, L. Qi, H. Ma, and J. Liu, "Surface Defects Detection of Paper Dish Based on Mask R-CNN," *Proceedings of the SPIE 10828, Third International Workshop on Pattern Recognition*, vol. 10828, 2018.

[22] H. Wen, L. Chen, T. Fu, Z. Yang, and Z. Yin, "Detecting the Surface Defects of the Magnetic- Tile Based on Improved YOLACT++," *Proceedings of the International Conference on Communication Technology* , Tianjin, China, Oct. 13–16, 2021, pp. 1097-1102.

[23] H. Fu, D. Meng, W. Wu, and Y. Wang. "Crack Segmentation Based on Improved YOLACT++ Algorithm," *Proceedings of the 2021 Chinese Intelligent Automation Conference*, Zhanjiang, China, Nov. 5–7, 2021, pp. 425-432.

[24] J. Liu, T. Liu, Y. Rong, R. Cao, and L. Tian, "Surface Defect Detection of Medium and Thick Plates Based on MASK-RCNN," *Proceedings of the Advances in Artificial Intelligence and Security, 2022 8th International Conference on Artificial Intelligence and Security*, Qinghai, China, July 15–20, 2022, vol. 1586, pp. 426-436.

[25] C. Huang, Y. Zhou, and X. Xie, "Intelligent Diagnosis of Concrete Defects Based on Improved Mask R-CNN," *Applied Sciences-basel*, vol.14, no. 10, p. 4148, 2024.

[26] H. Wang, M. Li, Z. Wang. "Rail Surface Defect Detection Based on Improved Mask R-CNN," *Computers & Electrical Engineering*, vol.102, pp. 108269-108269, 2022.

[27] K. He, G. Gkioxari, P. Dollar, and R. Girshick. "Mask R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, Oct. 22–29, 2017, pp. 2980–2988.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[29] J. Hu, L. Shen, G. Sun, "Squeeze-and-Excitation Networks," *Proceedings of the IEEE/CVF Conference on computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[30] D. Bolya, C.Zhou, F. Xiao, and Y. J. Lee. "YOLACT: Real-Time Instance Segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 9156-9165.

[31] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 8570-8578

[32] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting Objects by Locations," *Proceedings of the European Conference on Computer Vision*, Glasgow, UK, 2020, pp. 649-665.

[33] X.Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and Fast Instance Segmentation", *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 17721-17732.