IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 May 2024, accepted 4 June 2024, date of publication 11 June 2024, date of current version 18 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412425

RESEARCH ARTICLE

EHA-YOLOv5: An Efficient and Highly Accurate Improved YOLOv5 Model for Workshop Bearing Rail Defect Detection Application

JIYONG HU[1], HONGFEI YANG[2,3], (Member, IEEE), JIATANG HE[1], DONGXU BAI[3], (Member, IEEE), AND HONGDA CHEN[3,4]

[1]FAW-Volkswagen Automobile Company Ltd., Changchun, Jilin 130011, China
[2]School of Mechanical and Aerospace Engineering, Jilin University, Changchun 130061, China
[3]School of Instrumentation Science and Electrical Engineering, Jilin University, Changchun 130061, China
[4]Information Engineering College, Huzhou Normal University, Huzhou, Zhejiang 313000, China

Corresponding authors: Hongfei Yang (yanghf20@mails.hjlu.edu.cn) and Hongda Chen (chenhd20@mails.hjlu.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant 51875232 and Grant 51875233.

ABSTRACT Addressing the challenge of surface defect detection in load-bearing rails within auto-motive assembly workshops, which operate in complex environments and under long-term service, this paper proposes an innovative detection framework based on an improved YOLOv5 network. This framework, designed specifically for the unique challenges presented by load-bearing rails, integrates advanced machine vision and deep learning technologies. Initially, a Multi-Scale Pyramid Pooling (MSPP) module, incorporating the concept of residual stacking, is introduced to effectively enhance the extraction of complex features; Subsequently, the coordinate attention mechanism is optimized, leading to the development of a novel Spatial Coordinate Attention Mechanism (DAM), focused on detecting small-sized defects; Thereafter, a Dual Sampling Transition Module (DSTM) is applied to enhance information retention during the down-sampling process; Finally, the DBDAMN clustering algorithm is utilized to optimize anchor sizes, allowing for more precise adaptation to the diversity of defect sizes. These innovations significantly improve the accuracy of surface defect detection in load-bearing rails, particularly in identifying small defects, offering an effective means of preventing workshop safety incidents. The experimental results demonstrate that this method achieves 97.3% on AP50, marking a 4.2% improvement over the standard YOLOv5 model, thus indicating a significant performance enhancement. To validate the superiority of our model, a comparison with popular current models was conducted, achieving optimal values in recall rate, accuracy, and mAP, which were 91.4%, 92.6%, and 88.9%, respectively. Therefore, the proposed method meets the requirements for precision in rail defect detection.

INDEX TERMS YOLOv5, defect detection, dual attention mechanism, residual pyramid pooling model, DBDAMN clustering algorithm.

I. INTRODUCTION

Due to complex environments and extended service, various faults occur in the load-bearing rails of automotive assembly workshops during their operational lifespan. If not promptly addressed, these faults can lead to significant workshop accidents, resulting in immeasurable economic losses and severe risks to the safety of front-line installation workers, as exemplified by the 2018 collapse incident at the Chongqing Changan Automobile workshop [1]. Compared to standard

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

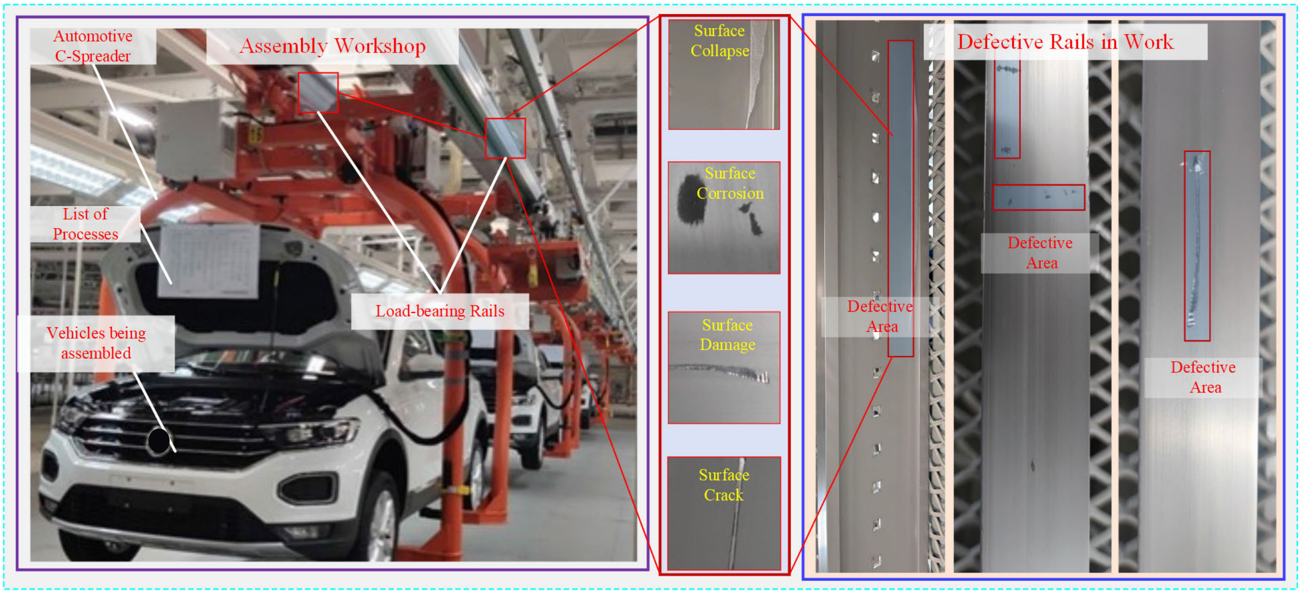


FIGURE 1. Rails in use in the workshop with typical cases of defects.

rails, workshop rails are subjected to higher relative motion speeds and greater vertical loads, thus facing a higher risk of failure. As shown in Fig 1, in automotive assembly workshops, the vehicle assembly process involves hoisting cars with C-type lifting devices, increasing the failure risk of the load-bearing rails. Prolonged use leads to overall deformation of the rails, causing surface material to crumble, and the complex workshop environment, often with oily liquids and gases, contributes to the corrosion of these heavy rails. The prolonged contact between the hoisting device's wheels and the rails, along with sudden stops and starts, significantly exacerbates surface defects on the rails. The most common surface defects include indentations, joint failures (breaks), peeling, rolling contact fatigue cracks, squeezing, lateral wear, delamination, and corrosion, leading to functional degradation of the rails. Therefore, accurate and reliable real-time detection of surface defects in assembly workshop rails, with early intervention to curb the development of defects, is a necessary measure to prevent major safety accidents in assembly workshops.

Currently, rail defect detection methods primarily include manual techniques such as tapping with tools, visual inspection, and laser magnetic leak-age, all of which are highly subjective [2], [3], [4], [5], [6], [7], [8]. The accuracy of these detection results is greatly influenced by the inspectors' work experience, technical expertise, physical and psychological state, and work environment. Accumulating experience in rail inspection takes considerable time, leading to experienced inspectors often being older and more prone to fatigue under high-intensity work conditions. Conversely, younger inspectors often lack sufficient experience, making it difficult to ensure the reliability of their inspection results. In workshops with loud noise and strong odors, the precision of laser and magnetic leakage detection equipment is severely compromised. The ongoing production processes pose significant threats to the safety of inspection personnel and can also impact the workshop's productivity, leading to substantial economic losses.

Thanks to advances in machine vision and deep learning, new solutions have emerged for the aforementioned problems [9], [10], [11], [12], [13], [14]. Ma et al. [15] introduced a novel one-shot unsupervised domain adaptation framework for the segmentation of rail surface defects. They introduced a shape-consistent style transfer module that performs pixel-level distribution alignment between training and test images. Xiao et al. [16] proposed a new small-sample defect classification method. Xiao et al. [16] developed a pixel-level defect segmentation approach. Xiao et al. [16] presented a dual-domain adaptive model for the detection of defects in automobile tires. Wang et al. [17] introduced an enhanced encoder-decoder network with hierarchical supervision. Xu et al. [18] proposed a design framework based on self-supervised representation learning. Zhang et al. [19] developed a multi-scale attention feature fusion module. Yang et al. [20] proposed a multi-level, end-to-end method for the rapid detection of surface defects on rails. Liu et al. [21] demonstrated the application of a pyramid feature convolutional neural network in the detection of surface defects on rails.

In this study, addressing the issue of defect detection in workshop load-bearing rails, we have developed an innovative and efficient defect detection framework. The proposed defect detection framework, based on an enhanced YOLOv5 network, addresses surface defect challenges in load-bearing rails within automotive assembly workshops. Its integration of advanced machine vision and deep learning technologies

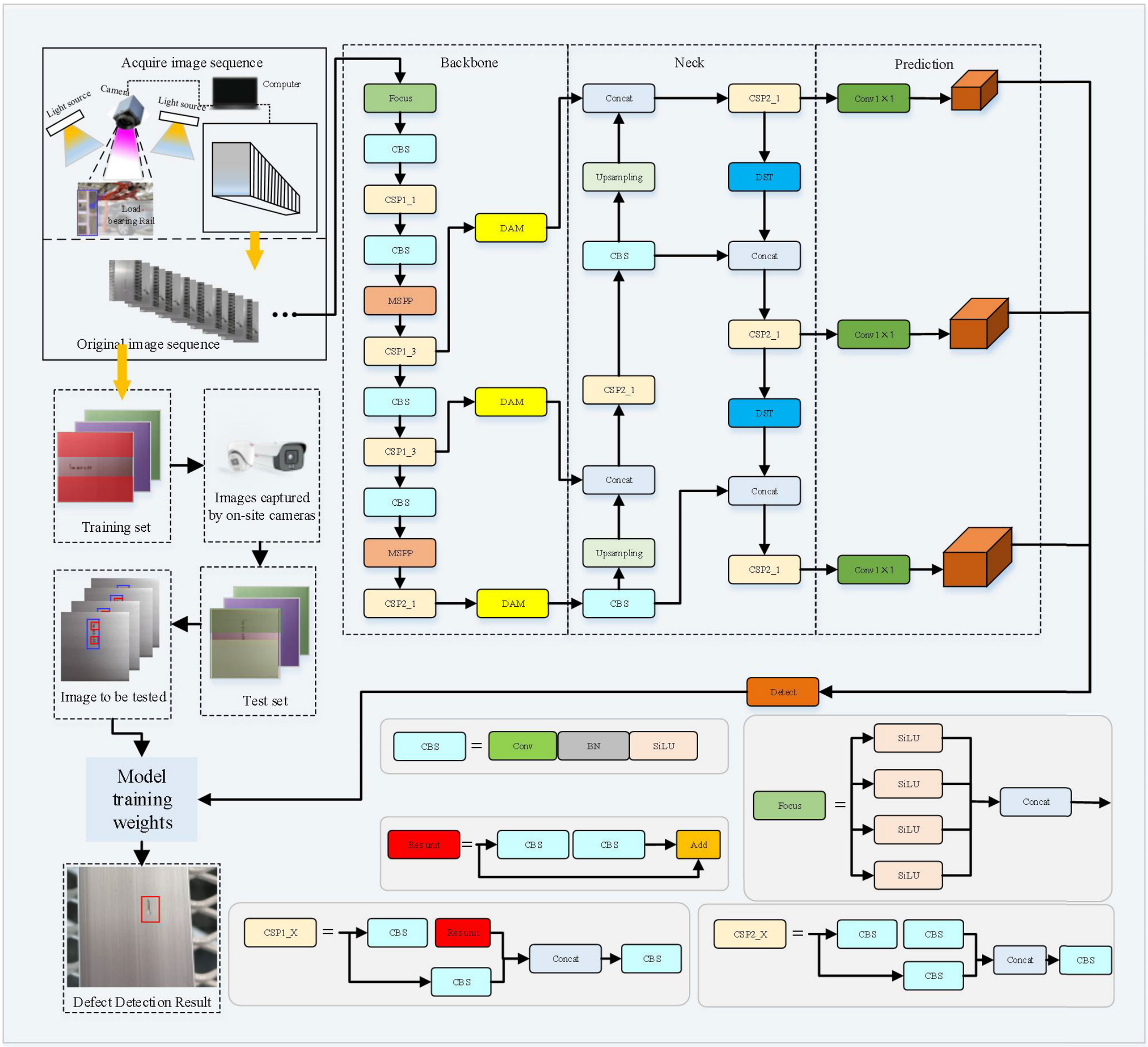


FIGURE 2. Brief architecture of the proposed EHA-YOLOv5 track defect identification network. The meanings of the various symbols are indicated at the top of the figure.

offers potential for generalization to diverse environments. Through transfer learning, environmental adaptation, and domain-specific tuning, the framework can be tailored to detect defects in various industrial settings. Real-time adaptation, collaborative learning, and data augmentation further enhance its versatility and robustness. Despite requiring adjustments, its core principles provide a strong foundation for scalable defect detection solutions beyond automotive assembly workshops. The main innovations and contributions of this paper can be summarized as follows:

1. This research improves the traditional. SPP module by adopting a residual stacking approach. The new M-SPP module, with its denser residual structure, enhances the extraction of detailed features by deepening the net-work architecture.

2. This study optimizes the coordinate attention mechanism, developing DAM, which can analyze the weight relationships between different pixels in space more meticulously.

3. In the Path Aggregation Network (PANet), this study introduces the DSTM module for down-sampling to preserve more critical information.

4. To more accurately adapt to the size distribution of defects on the sur-face of load-bearing rails, this study designs the DBDAMN clustering algorithm to optimize the size of Anchors.

The rest of this paper is organized as follows: Section II describes the methodological theory. Section III provides a comparative analysis of the experimental results. Section IV presents the conclusion.

II. METHODOLOGY AND DESIGN

To enhance the YOLO model's detection of surface defects in load-bearing rails, this study introduces a series of improvements to the YOLOv5 network [22]. Initially, by designing the MSPP module to replace the existing SPP module, the network's capability to handle defects of various sizes

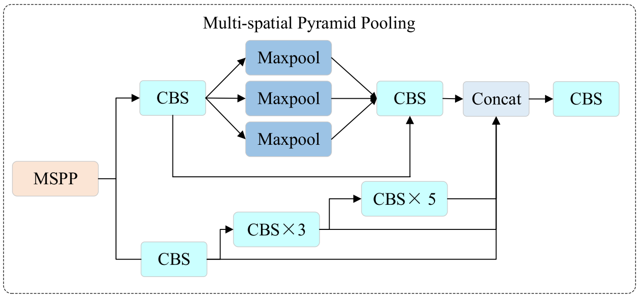


FIGURE 3. Schematic diagram of the specific structure of MSPP.

was improved, particularly enhancing target detection. Subsequently, by integrating the DAM attention mechanism, the network can focus more on key features, significantly improving the detection accuracy of small targets. Additionally, the design of the DSTM module to replace the down-sampling step further preserves important feature information, enhancing the network's feature extraction capability. Finally, to achieve more precise Anchor sizes, enabling the model to better adapt to the size distribution of rail surface defects, this paper utilizes the DBDAMN clustering algorithm to optimize Anchor settings, thereby improving overall accuracy. Fig2 displays the specific structure of the EHA-YOLOv5 network, fully demonstrating the synergistic role of design components in the overall architecture. Enhancing the network's feature extraction capability and the global attention mechanism is crucial for improving detection accuracy, significantly boosting the detection of minor defects in load-bearing rails.

A. BACKBONE NETWORK IMPROVEMENTS

In YOLOv5, the SPP (Spatial Pyramid Pooling) structure is integrated into the model. Although the SPP structure enables the extraction of richer spatial features, it may cause the model to overly rely on features of a specific scale. Particularly when there is a bias in the scale distribution of training data, such as most objects being of similar size, the model might learn more features of this specific scale. In such cases, the SPP layer may tend to extract features matching the main scale in the training data, overlooking information from other scales. This over-reliance could lead to a decrease in the model's generalization ability for inputs that appear different from the training data, even if they are of the same object type. To address these issues, this study proposes a method, MSPP, aimed at optimizing the performance of SPP by incorporating the concept of residual networks. This is particularly focused on enhancing the model's feature extraction capability and generalizability.

In the MSPP module, this paper draws on the design philosophy of residual networks to enhance the model's ability to process features of different scales. As shown in Fig 3, the MSPP module consists of two main parts: the upper SPP module and the lower CBS (Convolution, Batch Normalization, Activation) module. The CBS module includes a varying number of convolutional layers (namely 1, 3, and 5), each followed by a normalization layer and an activation function. The final output of the module is the stacked result of these four components. The upper CBS module is interconnected in two ways: the first involves connecting through a three-layer maxpool layer to the CBS module, and the second employs direct connections inspired by residual connections.

In the improved MSPP design, the SPP module is replaced with a denser residual structure, substituting the original SPP module in the YOLOv5 architecture. This design enables the model to combine inputs with extracted features, thereby obtaining deeper levels of information. The incorporation of the residual structure significantly improves the model's detection accuracy and also enhances its robustness in handling features of varying scales.

The residual unit can be expressed as follows:

$$y_l = h(x_l) + F(x_l, W_l), \tag{1}$$
$$x_l = f(y_l), \tag{2}$$

where x_l and x_{l+1} are the input and output of the l residual unit, respectively. F is the residual function, which represents the learned residuals, $h(x_l) = x_l$ represents the constant mapping, and f is the SiLU activation function. On the basis of the above equation, the learning characteristics from shallow l to deep L are obtained as follows:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i). \tag{3}$$

From the chain rule, the gradient of the reverse process can be obtained as follows:

$$\begin{aligned}\frac{\partial loss}{\partial x_l} &= \frac{\partial loss}{\partial x_L} \times \frac{\partial x_L}{\partial x_l}\\ &= \frac{\partial loss}{\partial x_L} \times (1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_i)).\end{aligned} \tag{4}$$

The first factor $\frac{\partial loss}{\partial x_L}$ represents the gradient of the loss function arriving at L; the 1 in parentheses indicates that the short-circuiting mechanism can propagate the gradient losslessly. The other residual gradient needs to pass through the layer with weights, and the gradient is not passed directly. The residual gradient is not all -1, and even if its value is relatively small, the presence of 1 does not cause the gradient to vanish. Thus, residual learning is easy.

From Equations (1) and (2), $h(x_l) = x_l$ denotes the constant mapping. We then have

$$y_{l+1} = h(x_{l+1}) + F(x_{l+1}, w_{l+1}), \tag{5}$$
$$y_{l+1} = x_{l+1} + F(x_{l+1}, w_{l+1}), \tag{6}$$
$$y_{l+1} = f(y_l) + F(f(y_l), w_{l+1}). \tag{7}$$

Suppose f is of asymmetric form, $f(y_l) = y_1$, and $f(y_l)$ is rewritten as $\hat{f}(y_l)$. We then have

$$x_{l+1} = x_l + F\left(\hat{f}(x_l), w_l\right). \tag{8}$$

An asymmetric activation function is applied, and the activation function first is used in the residual function part and

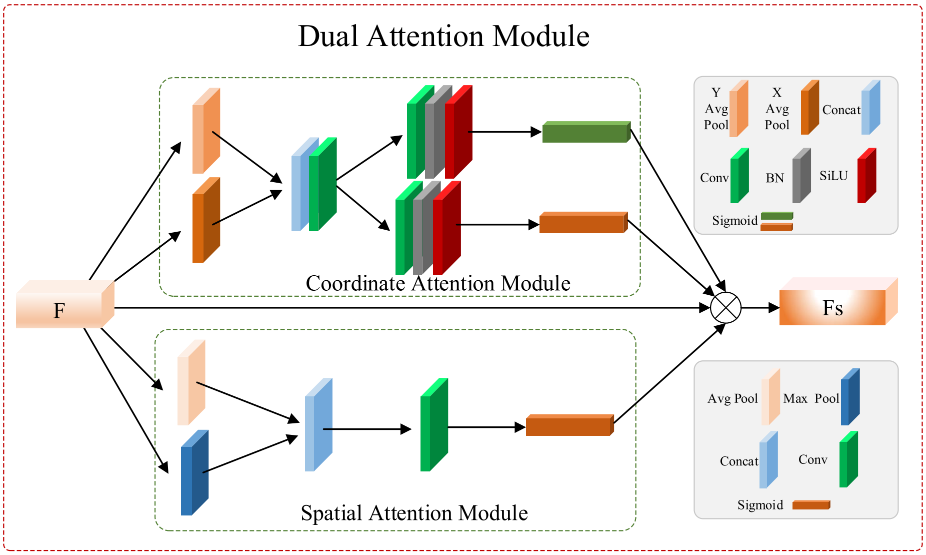


FIGURE 4. Schematic diagram of the specific structure of DAM.

in the calculation with weights *w*. Thus, the use of preactivation indicates that adjusting the position of SiLU and BN for preactivation causes the regularization effect and reduces overfitting, thereby resulting in higher accuracy.

The focus of the MSPP design is to address the shortcomings of the traditional SPP module in scenarios with uneven defect size distribution, particularly its over-reliance on specific scale features. By introducing a more complex residual structure, the MSPP module enhances the diversity of feature extraction, thereby reducing dependence on single-scale features. Furthermore, the incorporation of the residual structure also improves the model's adaptability and generalization capabilities when dealing with diverse inputs.

B. ATTENTION MECHANISM DESIGN

Due to the randomness, variety, and presence of many small defects in rail defect distribution, traditional Yolo detection methods experience missed and false detections. Therefore, incorporating the DAM mechanism into the EHA-YOLOv5 architecture is particularly important in this study. The integration of the DAM mechanism aims to enhance the target detection model's ability to recognize key features, especially in dealing with complex, small surface defects on rails.

The challenge of rail defect detection lies in accurately identifying and locating small, irregular defects on the rail surface. The DAM mechanism, by strengthening spatial and channel-level features, makes the model more sensitive to these small defects. This attention mechanism effectively distinguishes subtle differences between defect features and normal parts of the rail, thereby enhancing detection accuracy. Surface defects in load-bearing rails can present in various sizes and shapes. The DAM mechanism provides a more refined feature fusion capability within the YOLOv5 multi-scale feature extraction framework.

By effectively adjusting spatial and channel attention across different feature levels, the model's recognition ability under multiple scales is enhanced, which is crucial for identifying defects of various sizes and types. In practical applications, the detection environment for load-bearing rails may vary due to factors such as lighting and background interference. The DAM mechanism helps to improve the adaptability and generalization ability of the YOLOv5 model in these variable environments. By focusing more on key features, the model can better cope with environmental changes and maintain the stability of its detection performance. The framework demonstrates robustness to variations in environmental conditions, including lighting changes, camera angles, and surface textures, through several mechanisms. Firstly, robust feature extraction techniques, coupled with advanced deep learning architectures, enable the model to learn invariant representations of defects across different conditions. Secondly, data augmentation methods introduce variability during training, enhancing the model's ability to generalize to diverse environments. Additionally, post-processing techniques such as normalization and adaptive thresholding further stabilize detection performance under varying conditions. Finally, continuous monitoring and feedback mechanisms allow the framework to adapt to dynamic environmental changes over time, ensuring sustained robustness in real-world applications. Furthermore, the system itself is equipped with accompanying lighting devices, ensuring effective avoidance of excessively extreme environmental conditions beyond the adjustment range of the framework.

The specific structure of DAM, as shown in Fig 4, mainly consists of two attention mechanisms: CAM (Coordinate Attention Module) and SAM (Spatial Attention Module). The design of the CA attention mechanism aims to capture important information in both the width and height dimensions of rail images, effectively encoding precise defect location information. By performing global average pooling in both width and height directions to down-sample the defect feature map, the CA mechanism can obtain pooled defect feature maps in these two directions. These feature maps are then concatenated and processed through an activation function to obtain weights in these two spatial dimensions. The feature tensor in the middle of an arbitrary network can be expressed as follows.

$$X = [x_1, x_2, \cdots, x_{c-1}, x_c] \in R^{H\times W\times C} \tag{9}$$

Here, H and W represent the height and width of the rail feature map, respectively; C is the number of channels; x is the feature tensor of different channels.

$$Y = [Y_1, Y_2, \cdots, Y_{c-1}, Y_c] \in R^{H\times W\times C} \tag{10}$$

After the CA performs global average pooling on the input feature map, the output for height H and width W in channel C is

$$Z_c^h(h) = \frac{1}{W}\sum_{0\leq I<W} x_c(h, i) \tag{11}$$

$$Z_c^h(w) = \frac{1}{H}\sum_{0\leq I<H} x_c(j, w) \tag{12}$$

The obtained defect feature maps in width and height are stacked and fed into a 1 × 1 convolution module F1, reducing the dimension to the original C/r, then using the ReLU activation function δ to obtain the feature map, that is

$$f = \delta(F_1(\left[z^h, z^w\right])) \tag{13}$$

The processed rail defect feature map f is then convolved using a 1 × 1 convolution in both width and height to obtain the same number of channels as the initial feature map, resulting in defect feature maps F_w and F_h.Subsequently, the Sigmoid activation function σ is used to obtain weights in two dimensions: width and height.

$$\mathrm{g}^w = \sigma(F_w(f^w)) \tag{14}$$

$$\mathrm{g}^\mathrm{h} = \sigma(F_h(f^h)) \tag{15}$$

Finally, the original feature map is multiplied by the weights in both the width and height dimensions to obtain the final output.

$$y_c(i, j) = x_c(i, j) \times g_c^w(j) \tag{16}$$

The CA attention mechanism is capable of capturing positional information and the relationships between different channels. To further distinguish the weight relationships between different pixels in the spatial domain, focus more on areas of interest, and reduce the weights of non-essential defect-free areas, a spatial attention module is integrated on the basis of CA. Specifically, global average pooling and global max pooling are performed in the channel dimension, resulting in two rail defect feature maps $\mathrm{F}^\mathrm{s}_\mathrm{avg}$ and $\mathrm{F}^\mathrm{s}_{max}$, each with a single channel, which are then stacked to form an H × W × 2 feature map.

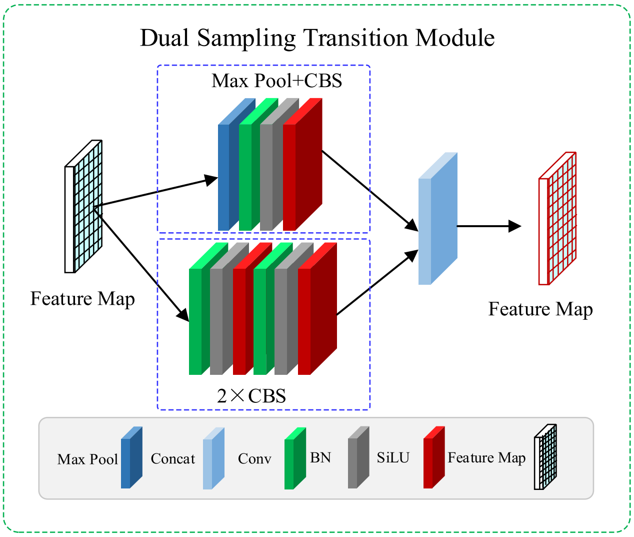


FIGURE 5. Schematic diagram of DSTM module structure.

$$m_s = \sigma(Conv^{7\times 7}([F^s_{avg}, F^s_{max}])) \tag{17}$$

A convolution with a 7 × 7 kernel is used to obtain an H × W × 1 feature map, and finally, the Sigmoid activation function is applied to obtain spatial attention weights, represented as follows:

$$w_s = x_i \times y_c \times m_s \tag{18}$$

The DAM attention mechanism aims to more effectively focus the model's "attention" on features that are crucial for the final detection task. It enhances important information and suppresses less relevant information by considering both spatial and channel features simultaneously. In this study, DAM is applied to three forward channels between the backbone network and PANet (Path Aggregation Network) to achieve attention adjustment across multiple feature layers at different levels of the network.

By applying DAM in the forward channels between the backbone network and PANet, precise attention adjustment can be achieved across multiple feature layers. This multi-level attention mechanism helps the network more effectively merge and utilize information from different layers. Particularly in object detection, this can aid the model in better distinguishing between targets and background, as well as identifying rail defects of various sizes and shapes.

C. DSTM MODULE

In YOLOv5, the PANet structure, serving as the "neck" of the network, effectively merges multi-scale features extracted from the backbone network. This paper aims to delve into the application of PANet in YOLOv5 and introduces a new down-sampling module called DSTM (Dual Sampling Transition Module) to further enhance the efficiency and accuracy of feature extraction.

Traditional down-sampling methods, while reducing the spatial dimensions of feature maps, also entail partial loss of information. To compensate for this deficiency, this study proposes the DSTM. DSTM combines two common down-sampling methods to maximally preserve important information during the down-sampling process.

Specifically, the design of DSTM includes two branches: the upper branch employs 2 × 2 max pooling, followed by a 1 × 1 convolution operation. This design aims to capture key spatial features through pooling, while 1 × 1 convolution is used to maintain channel consistency. The down branching uses two convolutional layers, 1 × 1 and 3 × 3 convolutional kernels, and the step size is set to 2.This branch is primarily responsible for extracting finer-grained spatial information. The results of both branches are stacked after their respective down-sampling operations, forming a composite feature map.

DSTM demonstrates significant advantages in enhancing the target detection performance of YOLOv5. By combining the two down-sampling methods, the DSTM not only preserves key spatial features, but also provides a richer feature representation. This is particularly important for detecting small-sized targets or targets within complex backgrounds. In practical applications, such as high-precision required scenarios like load-bearing rail defect detection, the application of DSTM significantly improves the model's detection accuracy and reliability.

D. DBDAMN RE-CLUSTERING ANCHORS

In the YOLOv5 algorithm, the size of Anchors must be predefined before training and prediction. In YOLOv3, 9 Anchors are predetermined through cluster analysis, with 3 set for each output scale. This resolves the issue of low recognition rates for small objects. Therefore, the detection accuracy is influenced by the initial Anchor settings, which need to be predefined in the algorithm. Due to the potential variations in size and distribution of rail defects, this paper improves the initialization method of Anchors to further enhance the detection accuracy of YOLOv5, particularly for defects of different sizes.

The effectiveness of the anchor optimization algorithm is closely tied to the representativeness of the training data. Anchors are predefined bounding box shapes used during object detection to predict object locations and sizes. When optimizing anchors, the algorithm relies on the distribution and characteristics of objects present in the training data. If the training data is representative of the target environment and contains a diverse range of object sizes, shapes, and aspect ratios, the anchor optimization algorithm can better adjust the anchor priors to match the distribution of objects in the data. This results in more accurate predictions during inference, as the model is better equipped to handle objects of various shapes and sizes present in real-world scenarios. However, if the training data is not representative or lacks diversity, the anchor optimization algorithm may struggle to adapt the anchor priors effectively. This can lead to suboptimal anchor configurations, resulting in reduced detection performance, particularly for objects that deviate significantly from the training data distribution.

Algorithm 1 DBDAMN Re-Clustering Anchors

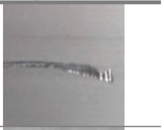
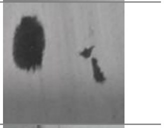
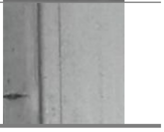
```
Input: Dataset D (a collection of defect points on
 the surface of the rail); Neighborhood radius Eps;
 Minimum number of points MinPts
Output: Defect clustering results
Begin algorithm
Function DBDAMN, (D,EPS,MinPts)
  Initialize all points' clustering labels as
 unclassified
  For point P in dataset D:
    If P is already classified or marked as noise:
      Initialize new cluster C
      Add P to cluster C
      Add each point in P's neighborhood N to
 the processing queue Q
      While processing queue Q is not empty:
        Take out a point $Q_p$
        If $Q_p$ is marked as noise:
          Add $Q_p$ to cluster C
        If $Q_p$ is already classified:

        Mark $Q_p$ as part of C
        Initialize $Q_p$'s neighborhood set Np
        If the number of neighborhood points
 $(N_p) >= $ MinPts.
        Add the points in $N_p$ to the processing
 queue Q
   Add cluster c to the clustering results
 Return clustering results
 End algorithm
```

The decision-making process of a deep learning model, particularly in the context of defect detection, is intricate and multifaceted. Initially, the model extracts pertinent features from input data, typically images, through convolutional layers. These features encapsulate patterns and attributes indicative of defects on load-bearing rails or other surfaces. Subsequently, the model employs classification techniques to discern the presence or absence of defects based on learned patterns and relationships between features and defect labels.

One critical aspect of the decision-making process involves thresholding. By applying a threshold to the model's output probability, a binary determination is made regarding the presence of a defect. Adjusting this threshold is pivotal, as it directly influences the model's sensitivity to false positives and false negatives. Finding the optimal threshold often entails striking a balance between minimizing false positives (incorrectly identifying defects where none exist) and false negatives (failing to detect actual defects).

Post-processing steps play a crucial role in refining the model's predictions and mitigating false positives or negatives. Techniques like non-maximum suppression or

TABLE 1. Dataset details.

Set	Types	Characteristics	Causes	Examples
L1	Crash	Surface material is partially cracked or missing and light reflection is uneven.	The presence of foreign material on the surface with the shop spreader extruded for a long period of time.	
L2	Damage	Surface material corroded and partially damaged.	Warm and humid environment around the assembly plant, presence of corrosive liquids and gases, etc.	
L3	Spot	Dirt, rust, and minor damage to the surface of the rails.	The presence of lubricating oil on surfaces, collisions between spreaders, etc.	

morphological operations are commonly employed for this purpose. Additionally, rigorous evaluation using performance metrics such as precision, recall, and F1-score provides insights into the trade-offs between false positives and false negatives. These metrics guide adjustments to the model's architecture, training data, or hyperparameters to enhance overall performance.

Furthermore, incorporating domain knowledge is invaluable. Experts in the field can offer insights into specific contexts or scenarios where false positives or false negatives may be more prevalent. By leveraging domain expertise, the model can be fine-tuned to better navigate these challenges.

In essence, the decision-making process of a deep learning model in defect detection encompasses feature extraction, classification, thresholding, post-processing, and integration of domain knowledge. Through iterative refinement guided by evaluation metrics and domain expertise, the model strives to strike an optimal balance between minimizing false positives and false negatives while maximizing overall performance and reliability in detecting surface defects.

Although the traditional K-means clustering algorithm [23] performs well in adapting to the distribution of target sizes in datasets, its random selection of initial positions may lead to variability and instability in clustering results. To address this issue, we propose the use of an improved clustering algorithm, DBDAMN (Density-Based Spatial Clustering of Applications with Noise).The DBDAMN algorithm does not rely on the choice of initial points and is better at handling outliers and noise in the data.

In the YOLOv5 model targeting surface defects of load-bearing rails, the DBDAMN algorithm is used to automatically determine Anchor sizes to better adapt to the size distribution of defects in the dataset. The DBDAMN algorithm determines clustering centers by analyzing the density connectivity of samples, making it more suitable for industrial image data with complex distribution characteristics.

Steps of the DBDAMN Re-Clustering Anchors Algorithm. Step 1: Select an appropriate neighborhood radius (Eps) and minimum number of points (MinPts). Step 2: For each defect point in the dataset, calculate the number of other defect points within its neighborhood. Step 3: Mark non-core points in the neighborhood that include core points as boundary points. Step 4: Group the core points and all points reachable by density into the same cluster. Step 5: After completing the clustering, analyze each cluster. The pseudo-code is shown below.

In consideration of load-bearing rail defect detection, the selection of Eps and MinPts must account for the significant variability in the size and distribution of rail defects. High-density defect areas may exist on the rails, and the DBDAMN algorithm can effectively process these areas to accurately identify the actual defects.

To match the needs of the YOLOv5 multi-scale detection heads, the number of clusters (k-value) is set to 9. Through multiple iterations and optimization processes of the DBDAMN algorithm, 9 different sizes of Anchors are finally determined. These Anchors more accurately reflect the actual size distribution of surface defects on load-bearing rails. The application of the DBDAMN algorithm in detecting surface defects of load-bearing rails enhances the accuracy of defect identification and increases the detection capability of complex defect patterns. With appropriate parameter settings and detailed analysis of clustering results, the DBDAMN algorithm can identify key defect areas in rail defect data.

III. EXPERIMENTATION AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENTS AND DATA SETS

The assembly workshop load-bearing rail image dataset used in this paper was extracted from videos, with the camera lens directly facing the rail. The different service times of the rails result in various types of noise in the corresponding image samples. The noise contained in the image samples is part of the rail image dataset, which includes 1036 samples. The dataset was expanded to 4,500 images using data augmentation methods such as rotation, mirroring, flipping, scaling, adding noise, and label smoothing. Various methods exist for acquiring training data, including field data

TABLE 2. Comparison of ablation experiment results.

DBDAMN	MSPP	DAM	DSTM	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	×	×	×	63.3	93.1	67.5	38.8	61.3	68.3
√	×	×	×	64.5	94.5	68.9	38.9	62.1	68.5
×	√	×	×	63.6	93.8	67.3	39.4	62.4	69.6
×	×	√	×	66.1	96.4	68.5	41.3	62.6	69.3
×	×	×	√	65.5	95.6	67.2	38.7	62.2	69.1
√	√	√	√	**66.8**	**97.3**	**70.3**	**41.6**	**64.3**	**71.2**

TABLE 3. Comparison of DAM using different pooling results.

YOLOv5	Parameters（M）	AP_S(%)	AP_M(%)	AP_L(%)
+CA(GAP)	180.5	40.8	60.8	68.3
+CA(GMP)	180.5	41.2	62.0	68.9
+DAM(OURS)	180.5	41.3	62.6	69.3

collection using cameras or sensors, manual or semi-automated annotation, utilizing publicly available datasets, generating synthetic data via computer graphics, employing transfer learning, augmenting data through transformations, and utilizing crowdsourcing platforms like Amazon Mechanical Turk. These methods enable the acquisition of diverse and high-quality training data, crucial for effective model training across different domains. Combining these approaches ensures comprehensive support for model training, enhancing its robustness and generalization capabilities. The dataset used during training is in VOC format, and the annotated and enhanced dataset is divided into training and test sets in a 9:1 ratio. As shown in Table 1, L1 includes rail surface material breakage and uneven surface reflection; L2 involves rail surface deterioration and material damage due to prolonged service; L3 contains numerous small damages and noise on the rail surface. The paper discusses various iterations and learning rates of the training network to achieve superior segmentation capability.

The challenges were faced in annotating comprehensive defect annotations include subjectivity, complexity, variability, time-consuming, expertise and labeling guidelines. Addressing these challenges often involves implementing standardized annotation protocols, providing annotator training, leveraging automation and semi-automation tools, and conducting rigorous quality control measures.

For computational Requirements, the framework's computational demands primarily stem from the deep learning model's inference process. The enhanced YOLOv5 network is typically computationally intensive, requiring powerful hardware such as GPUs or specialized accelerators for efficient processing. Additionally, preprocessing steps, feature extraction, and post-processing also contribute to computational overhead. Optimizations such as model compression and efficient hardware utilization are crucial for achieving real-time performance. Rigorous testing across diverse industrial environments is necessary to validate its effectiveness. With proper optimization and validation, the framework shows promise for practical implementation in automotive assembly workshops and similar industrial settings, offering improved defect detection accuracy while mitigating false alarms.

Our network experiments were conducted on a computer with an Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz, Tesla V100s GPU, 256GB RAM. The initial learning rate of the paper is 0.01, with a minimum learning rate of 0.0001, using an SGD optimizer, momentum of 0.937, weight decay rate of 0.00005, mosaic data augmentation method, 200 iterations, and training with the YOLOv5l model. Each experiment was conducted 10 times to obtain average results.

B. ABLATION EXPERIMENT

To assess the effectiveness of various improvements to the YOLOv5 network in detecting surface defects of load-bearing rails, we employed the Average Precision (AP) based on Intersection over Union (IoU), including overall AP and its performance at different IoU thresholds (AP50, AP75) and object sizes (APS, APM, APL) as key evaluation metrics. Through a series of experiments on a unified dataset, we found that the Anchor optimized by the DBDAMN clustering method increased the AP and its performance at different IoU thresholds by 1.2%, 1.4%, and 1.4%, respectively. This improvement is mainly attributed to the optimized Anchor more accurately matching the size characteristics of surface defects on the rails. In the application of the MSPP module, the detection performance for small (APS), medium (APM), and large (APL) targets improved by 0.6%, 1.1%, and 1.3%, respectively, demonstrating enhanced detection effectiveness for larger targets.

Additionally, the introduced DAM attention mechanism significantly enhanced performance in the network's three output layers, with AP, AP50, and AP75 increasing by 1.5%, 1.8%, and 0.8%, respectively. For targets of different sizes, the improvements in APS, APM, and APL were 2.8%, 0.9%, and 0.8%, respectively, with the most notable enhancement in detecting small targets (APS). When the DSTM module was used in the network's down-sampling process, all evaluation metrics showed improvements, indicating the module's effectiveness in feature fusion. After comprehensively applying the DBDAMN clustering algorithm, MSPP, DAM, and

TABLE 4. Comparison of different attention mechanisms.

YOLOv5	Parameters（M）	AP_S(%)	AP_M(%)	AP_L(%)
+SE	180.2	38.9	61.4	68.8
+CBAM	180.7	39.5	62.5	69.1
+CA	180.4	39.9	61.3	68.2
+DAM	180.5	**41.3**	**62.6**	**69.3**

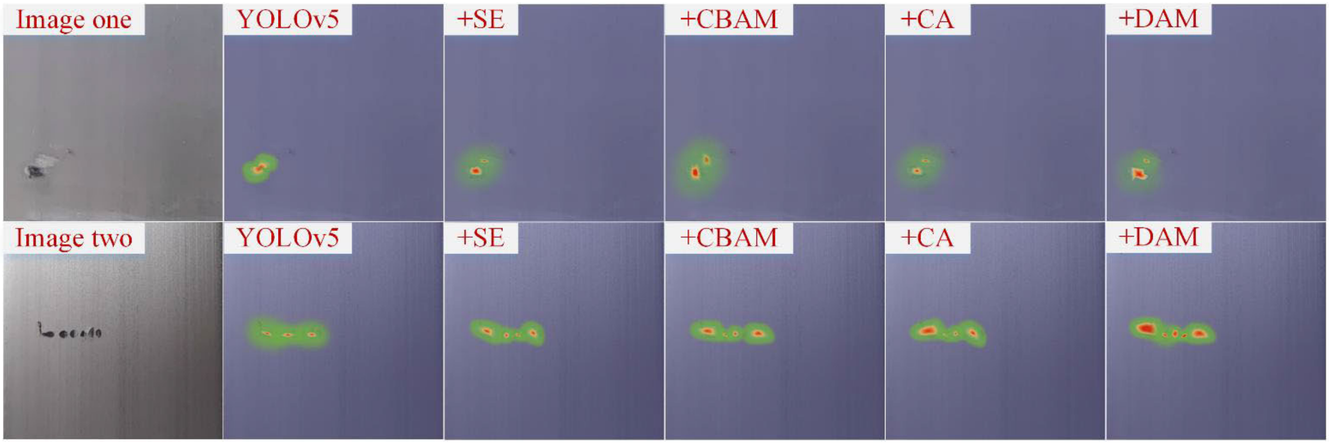


FIGURE 6. Schematic diagram of the results of the comparison of the heat maps of the different attention mechanisms.

DSTM, the increases in AP, AP50, and AP75 were 2.5%, 4.2%, and 2.8% respectively, and APS significantly increased from 38.8% to 41.6% (an increase of 2.8%), with APM and APL also improving by 3% and 2.9%, respectively.

These results show that by incorporating these improved strategies in the YOLOv5 network, the network not only strengthens the feature extraction capability of the network for load-bearing rail defects, but also improves the ability of the network to capture global information, thus significantly enhancing the overall accuracy of rail defect detection.

C. COMPARISON OF ATTENTION MECHANISMS

To assess the contribution of the improved DAM in enhancing network detection accuracy in this study, we compared the effects of using GAP alone, GMP alone, and the improved DAM method in the CA (Coordinate Attention) mechanism [24]. According to the experimental results, the model's parameter count remains essentially unchanged under these three methods. The results show that when only GAP is applied, the DAM method proposed in this study improved the performance in APS and APM by 0.5% and 1.8% respectively, and also increased APL performance by 1.0%. In the scenario of using GMP, the performance of APS and APM decreased by 1.1% and 0.6%, respectively. Although the improvement in APL was not significant, the DAM model also improved by 0.4%. These findings indicate that simultaneously applying GAP and GMP in the spatial attention mechanism can achieve the best results without increasing the model's parameter count. Moreover, these results also highlight the superiority of the improved DAM method we proposed.

To demonstrate the benefits of the improved DAM in enhancing the accuracy of detecting surface defects in load-bearing rails, we conducted a comparative analysis with several other popular attention mechanisms. Specifically, we selected SE (Squeeze-and-Excitation) [25], CBAM (Convolutional Block Attention Module) [26], CA (Coordinate Attention), and SCA (Spatial Channel Attention) proposed in this study for experimental comparison, with detailed results shown in Table 4. In the experiments, we focused on examining the differences in parameter quantity and detection effectiveness of these models.

The experimental data shows that in terms of model parameters, DAM has approximately 0.2M fewer parameters compared to the CBAM model. Under the YOLOv5 framework, the SE, CBAM, CA, and DAM models achieved improvements of 0.1%, 0.7%, 1.1%, and 2.8% respectively on the APS evaluation metric. Notably, DAM and CBAM performed best on the APM metric, with increases of 1.2% and 1.3% respectively, demonstrating strong performance in detecting medium-sized objects. Several different attention networks also achieved improvements on the APL evaluation metric, with the DAM model standing out with a performance of 69.3%.

Overall, these results demonstrate the superiority of DAM in detecting surface defects in load-bearing rails, especially in the detection of small-sized targets. The DAM mechanism, while minimally increasing the parameter count, effectively enhances the network's ability to recognize minute defects,

TABLE 5. Model comparison.

Model	Recall	Precision	mAP
Faster R-cnn	80.8	87.2	83.6
YOLOv4	82.6	86,8	81.3
YOLOv5	**89.0**	**88.2**	**87.8**
YOLOv7	86.8	87.5	86.2
YOLOv5（ours）	**91.4**	**92.6**	**88.9**

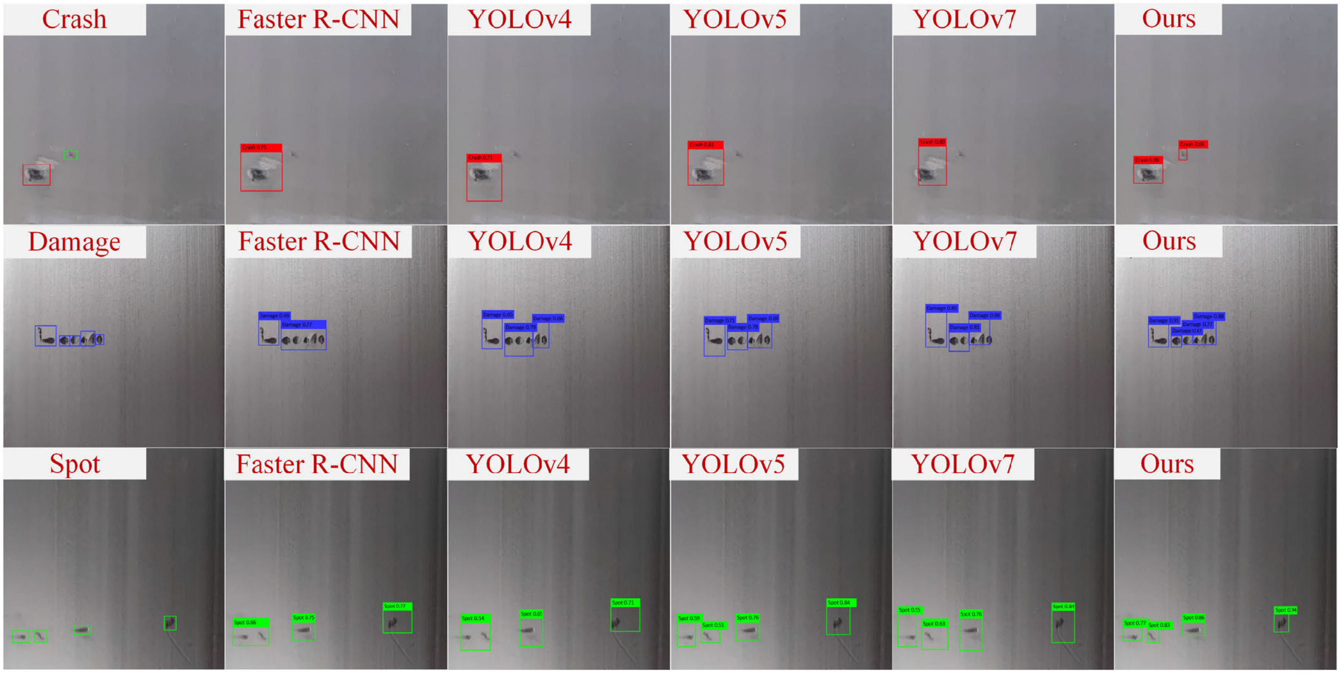


FIGURE 7. Schematic of the detection results of different recognition models.

achieving higher detection accuracy compared to SE, CBAM, and CA networks. This finding emphasizes that in designing efficient attention mechanisms, it is important to consider both parameter efficiency and the improvement of detection performance.

To visually demonstrate the application effectiveness of our improved DAM in detecting surface defects of load-bearing rails, we employed CAM technology to analyze the heatmap outputs of YOLOv5 and different attention mechanisms, as exemplified in Fig6. This analysis aims to assess the efficacy of each attention mechanism in locating small-sized defects.

The experimental results revealed that in the standard YOLOv5 network, the detection of certain minute defects is not very significant, indicating insufficient focus on the target. The performance improved with the introduction of different attention mechanisms. Specifically, the SE mechanism, although helpful in locating some defects, did not show clear coverage in the heatmaps, leading to some degree of missed detections. CBAM performed better in this regard, capturing defect locations more accurately, thanks to its simultaneous focus on channel and spatial information processing. However, the added CA mechanism still had issues with missed detections, particularly in accurately locating some small targets.

In contrast, the DAM model proposed in this paper shows more pronounced heatmap coverage at defect locations, significantly impacting the prediction results. This is particularly evident in detecting small defects such as multiple surface breakages and scratches. For larger defects such as material indentations, the DAM model's effectiveness is similar to before improvement, but there is a significant enhancement in detecting small defects. This demonstrates that the DAM model is more effective in differentiating the weight relationships between pixels in the spatial domain, focusing more on areas of interest. This also reduces attention to non-critical areas, thereby enhancing the overall detection accuracy and reliability. This improvement is particularly important when detecting surfaces of load-bearing rails with multiple minute defects. These minute defects are prone to be missed in the original YOLOv5 model, thus affecting the overall detection performance. Therefore, the algorithmic improvements in this paper are mainly focused on optimizing these easily missed minute defects to improve detection metrics.

D. MULTI-MODEL COMPARE

In this study, we adopted a comprehensive comparative experimental method to accurately assess the performance of the model proposed in this paper for detecting surface defects in load-bearing rails. For this purpose, we trained multiple network models on the same dataset, including the classic two-stage network Faster R-CNN [27], and the YOLO series: YOLOv4 [28], YOLOv5, and YOLOv7 [29], and conducted a detailed comparative analysis. The experimental results are summarized in Table 5, thereby providing a comprehensive perspective on performance evaluation.

The experimental results show that, in terms of the Recall metric, the improved YOLOv5 model proposed in this paper surpassed Faster R-CNN, YOLOv4, and YOLOv7, achieving respective improvements of 8.2%, 6.4%, and 2.2%. This achievement is mainly attributed to the efficiency of the improved YOLOv5 model in capturing small and complex defects, particularly showing better recognition capability in high-density and complex background conditions. Additionally, compared to the standard YOLOv5 model, our model improved the Recall from 89.0% to 91.4%, further proving the effectiveness of the improvements we made.

In terms of the Precision metric, the YOLOv5 network also excels, surpassing Faster R-CNN, YOLOv4, and YOLOv7. The improvements were 1.0%, 1.4%, and 0.7%, respectively. The model in this paper improved from 88.2% to 92.6% on this basis. This indicates that the model proposed in the paper has significant advantages in accurately identifying defects in load-bearing rails, particularly in reducing false positives and improving detection accuracy.

In terms of the comprehensive performance metric mAP, the model in this paper also shows significant advantages. Compared to Faster R-CNN, YOLOv4, and YOLOv7, it achieved improvements of 4.2%, 6.5%, and 1.6%, respectively. And compared to the standard YOLOv5 model, it increased from 87.8% to 88.9%. Achievement is mainly due to the improvements in the paper's model in handling multi-scale targets and enhancing overall detection accuracy.

In summary, this study comprehensively confirms the significant superiority of the improved YOLOv5 model proposed in this paper in the detection of surface defects on load-bearing rails through comparative analysis. These improvements not only enhance the model's ability to recognize small and complex targets, but also excel in reducing false detections and improving overall detection accuracy.

E. MULTI-MODEL VISUALIZATION COMPARISON

To demonstrate the superiority of the model in this paper, a detailed visual analysis was conducted, as shown in Fig7. There are two defective targets in the Crash sample, a large target (in the red box in the first row and column) and a small target (in the green box in the first column of the first row). The detection results show that all recognition models can identify the large target well. However, in the process of identifying small defects, other models missed the small defects(Faster R-CNN, YOLOv4, YOLOv5, YOLOv7). Due to the good sensitivity of the DAM mechanism proposed in this paper to small defects, only the model in this paper can detect them well. In terms of Anchor localization, YOLOv4, YOLOv5 and YOLOv7 have all experienced localization, and only the Faster R-CNN model and the model Anchor localization in this paper are more accurate. In terms of recognition accuracy, the other models are below 83% in recognizing large targets and only this paper's model achieves 89%, and only this paper's model detects in small target recognition, while all the other models show missed detection phenomenon. In the output results, we observed a certain degree of misdetection, which can lead to wrong judgment in industrial inspection applications. For this reason, the ability to recognize the misdetection cases is significantly improved by the improved network, which effectively avoids such misjudgments.

In the Damage sample, there are continuous medium-sized defects. In terms of Anchor positioning, the Faster R-CNN, YOLOv4, and YOLOv5 models all exhibited over-positioning of anchor boxes, while the YOLOv7 model showed under-positioning during the localization process. Meanwhile, all four models failed to precisely locate each medium-sized defect. Due to the attention mechanism and improved Anchor method proposed in this paper, which have good perceptual recognition for each defect, only the model presented in this paper accurately positioned the anchor boxes and identified the defects. In terms of recognition accuracy, the model in this paper achieved a defect detection accuracy above 75%, with the highest reaching up to 91%.

In the Spot sample with discrete medium-sized defects, Faster R-CNN and YOLOv4 incorrectly positioned the first two defect Anchors in a continuous manner during individual localization. YOLOv5, YOLOv7, and the model in this paper were all able to individually frame the defect locations effectively with their Anchor boxes. However, in terms of localization accuracy, the Anchor boxes of the model in this paper were the most accurate, without any over-positioning or under-positioning occurrences. In terms of recognition accuracy, all four models (Faster R-CNN, YOLOv4, YOLOv5, and YOLOv7) have less than 70% recognition accuracy for the first two defects, but the models in this paper are 77% and 83%, respectively. In the last two defects this paper's model also achieves the best scores, 86% and 94% respectively, while all other models are below 84%.

IV. EXPERIMENTATION AND ANALYSIS

This study aims to address the challenges of detecting surface defects on load-bearing rails in automotive assembly workshops in industrial applications, particularly focusing on the issues of low detection accuracy, high miss and false alarm rates in existing methods. To this end, the paper combines machine vision technology and deep learning methods to propose an assembly workshop load-bearing rail defect detection algorithm based on an improved YOLOv5. This algorithm has achieved significant success in improving the accuracy of surface defect detection and reducing the miss rate of small target defects. By employing the DBDAMN algorithm to recluster the initial Anchors of YOLOv5, the computational burden on the network was effectively reduced. Additionally, by integrating the MSPP module into the backbone structure of the network, the stacking and fusion of multi-scale features were achieved. This not only deepened the network structure but also improved accuracy and effectively mitigated the problem of gradient vanishing. In addition, this paper proposes the DAM mechanism and integrates it into the three forward channels between the backbone network and the PANet, which enables the network to focus more on small defects on the load-bearing rails, and effectively solves the problem of leakage detection of defects with small targets. Meanwhile, the application of the DSTM module in the down-sampling process of the feature map enables the network to capture more useful features, thus improving the overall detection accuracy.

Trade-offs between detection speed and accuracy are common in computer vision tasks. Increasing speed often requires simplifying models or reducing the complexity of computations, which can compromise accuracy. Conversely, improving accuracy may involve employing more complex models or performing extensive computations, leading to slower detection. The framework addresses these trade-offs by optimizing model architectures, leveraging efficient inference techniques, and implementing hardware acceleration where feasible. By striking a balance between speed and accuracy through careful design and optimization, the framework aims to achieve fast and reliable defect detection in real-world industrial environments.

Applying the method proposed in this paper, an AP50 of up to 97.3% detection accuracy is achieved in terms of accurate detection of rail surface defects, which is a 4.2% improvement compared to the traditional YOLOv5 model. Future research will continue to explore and incorporate improved methods to further enhance the accuracy and robustness of surface defect detection on load-bearing rails. When deploying the automated defect detection systems designed in this paper in industrial environments, we also took into account ethical considerations such as Worker Privacy, Job Displacement, Bias and Fairness, Safety and Reliability, Data Security and Integrity, and Environmental Impact.

REFERENCES

[1] (2018). [Online]. Available: http://www.cneb.gov.cn/2018/04/13/ARTI1523601215298234.shtml

[2] M. A. Machado, L. F. S. G. Rosado, N. A. M. Mendes, R. M. M. Miranda, and T. J. G. dos Santos, "New directions for inline inspection of automobile laser welds using non-destructive testing," *Int. J. Adv. Manuf. Technol.*, vol. 118, nos. 3–4, pp. 1183–1195, Jan. 2022.

[3] Y. Zeng, X. Wang, X. Qin, L. Hua, and M. Xu, "Laser ultrasonic inspection of a Wire+Arc additive manufactured (WAAM) sample with artificial defects," *Ultrasonics*, vol. 110, no. 3, pp. 106273–106284, 2021.

[4] W. Xu, J. Zhang, X. Li, S. Yuan, G. Ma, Z. Xue, X. Jing, and J. Cao, "Intelligent denoise laser ultrasonic imaging for inspection of selective laser melting components with rough surface," *NDT E Int.*, vol. 125, Jan. 2022, Art. no. 102548.

[5] Y. Long, S. Huang, L. Peng, S. Wang, and W. Zhao, "A characteristic approximation approach to defect opening profile recognition in magnetic flux leakage detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[6] J. Tang, R. Wang, B. Liu, and Y. Kang, "A novel magnetic flux leakage method based on the ferromagnetic lift-off layer with through groove," *Sens. Actuators A, Phys.*, vol. 332, Dec. 2021, Art. no. 113091.

[7] Z. Jin, M. A. I. Mohd Noor Sam, M. Oogane, and Y. Ando, "Serial MTJ-based TMR sensors in bridge configuration for detection of fractured steel bar in magnetic flux leakage testing," *Sensors*, vol. 21, no. 2, p. 668, Jan. 2021.

[8] J. Wu, W. Wu, E. Li, and Y. Kang, "Magnetic flux leakage course of inner defects and its detectable depth," *Chin. J. Mech. Eng.*, vol. 34, no. 1, pp. 1–11, Dec. 2021.

[9] J. Wang, P. Fu, and R. X. Gao, "Machine vision intelligence for product defect inspection based on deep learning and Hough transform," *J. Manuf. Syst.*, vol. 51, pp. 52–60, Apr. 2019.

[10] L. Zhu, P. Spachos, E. Pensini, and K. N. Plataniotis, "Deep learning and machine vision for food processing: A survey," *Current Res. Food Sci.*, vol. 4, pp. 233–249, Jun. 2021.

[11] D. Li and L. Du, "Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 4077–4116, Jun. 2022.

[12] A. Nasirahmadi, B. Sturm, S. Edwards, K.-H. Jeppsson, A.-C. Olsson, S. Müller, and O. Hensel, "Deep learning and machine vision approaches for posture detection of individual pigs," *Sensors*, vol. 19, no. 17, p. 3738, Aug. 2019.

[13] N. Sharma, R. Sharma, and N. Jindal, "Machine learning and deep learning applications-a vision," *Global Transitions Proc.*, vol. 2, no. 1, pp. 24–28, Sep. 2021.

[14] X. Lin, X. Wang, and L. Li, "Intelligent detection of edge inconsistency for mechanical workpiece by machine vision with deep learning and variable geometry model," *Int. J. Speech Technol.*, vol. 50, no. 7, pp. 2105–2119, Jul. 2020.

[15] S. Ma, K. Song, M. Niu, H. Tian, Y. Wang, and Y. Yan, "Shape consistent one-shot unsupervised domain adaptation for rail surface defect segmentation," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 1–12, Sep. 2022, doi: 10.1109/TII.2022.3233654.

[16] W. Xiao, K. Song, J. Liu, and Y. Yan, "Graph embedding and optimal transport for few-shot classification of metal surface defect," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022.

[17] J. Wang, G. Xu, C. Li, G. Gao, and Q. Wu, "SDDet: An enhanced encoder–decoder network with hierarchical supervision for surface defect detection," *IEEE Sensors J.*, vol. 23, no. 3, pp. 2651–2662, Feb. 2023, doi: 10.1109/JSEN.2022.3229031.

[18] Y. Xu, H. Wang, Z. Liu, and M. Zuo, "Self-supervised defect representation learning for label-limited rail surface defect detection," *IEEE Sensors J.*, vol. 23, no. 23, pp. 29235–29246, Dec. 2023, doi: 10.1109/jsen.2023.3324668.

[19] Z. Zhang, W. Wang, and X. Tian, "Semantic segmentation of metal surface defects and corresponding strategies," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023, doi: 10.1109/TIM.2023.3282301.

[20] H. Yang, Y. Wang, J. Hu, J. He, Z. Yao, and Q. Bi, "Deep learning and machine vision-based inspection of rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2021.3138498.

[21] Y. Liu, H. Xiao, J. Xu, and J. Zhao, "A rail surface defect detection method based on pyramid feature and lightweight convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022.

[22] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOV5: Improved YOLOV5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[23] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020.

[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] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[26] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[27] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 1440–1448.

[28] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOV4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[29] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

JIYONG HU received the bachelor's and master's degrees in engineering from Jilin University, in 2009 and 2012, respectively. He is currently a Planning Engineer with FAW-Volkswagen Changchun Branch. His main research interests include intelligent automotive assembly and conveyor systems and machine vision intelligent inspection.

HONGFEI YANG (Member, IEEE) received the master's degree in mechanical engineering from Jilin University, Changchun, China, in 2020, and the Ph.D. degree from the School of Instrumentation Science and Electrical Engineering, Jilin University, in September 2023.

He is currently a directly appointed an Associate Professor. His research interests include intelligent magnetometric sensing instrumentation, industrial defect detection, and environmental identification for engineering vehicles.

JIATANG HE received the B.Sc. degree in engineering and automation from Wismar University Germany Electrical, in 2007.

He is currently a Supervisor Planning Engineer with FAW-VW. His main interest includes conveyor system of automobile assembly shop.

DONGXU BAI (Member, IEEE) was born in Jilin, China, in 1998. He received the B.S. degree from Jilin University, in 2020, where he is currently pursuing the Ph.D. degree with the College of Instrumentation and Electrical Engineering.

His current research interests include geological instrumentation, intelligent geomagnetic sensing, natural disaster magnetic anomaly monitoring, high precision bell-bloom, and intelligent miniature magnetic sensor.

HONGDA CHEN received the Ph.D. degree from Jilin University, in 2023.

He is currently a Lecturer with the School of Information Engineering, Huzhou Normal University. His research interests include infrared spectroscopy detection and Photoacoustic spectroscopy gas detection.