IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 March 2024, accepted 15 March 2024, date of publication 25 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380009

RESEARCH ARTICLE

An Improved YOLOv8 Algorithm for Rail Surface Defect Detection

YAN WANG[1], KEHUA ZHANG[2,3], LING WANG[1], AND LINTONG WU[2]

[1]School of Engineering, Zhejiang Normal University, Jinhua 321004, China
[2]Key Laboratory of Urban Rail Transit Intelligent Operation and Maintenance Technology and Equipment of Zhejiang Province, Zhejiang Normal University, Jinhua 321004, China
[3]Intelligent Manufacturing Research Institute of Jinhua, Jinhua 321002, China

Corresponding author: Kehua Zhang (zhangkh207@zjnu.edu.cn)

This work was supported by the Science and Technology Planning Project of Jinhua under Grant 2022-1-096.

ABSTRACT To tackle the issues raised by detecting small targets and densely occluded targets in railroad track surface defect detection, we present an algorithm for detecting defects on railroad tracks based on the YOLOv8 model. Firstly, we enhance the model's attention towards small and medium-sized targets by substituting replacing the original convolution with the SPD-Conv building block in the backbone network of YOLOv8n, while preserving the original network structure. Secondly, we integrate the EMA attention mechanism module into the neck component, allowing the model to leverage information from different layers of features and improve feature representation capabilities. Lastly, we substitute the original C-IOU with the Focal-SIoU loss function in YOLOv8, which adjusts the weights of positive and negative samples to penalize difficult-to-classify samples more heavily. This enhancement improves the model's capability to accurately recognize challenging samples and ensures that the network allocates greater attention to each target instance, resulting in improved performance and effectiveness of the model. The experimental results reveal notable advancements in precision, recall, and average accuracy attained by our enhanced algorithm. Compared to the original YOLOv8n model, our enhanced algorithm demonstrates remarkable precision, recall, and average accuracy of 93.9%, 93.7%, and 94.1%, respectively. These improvements amount to 3.6%, 5.0%, and 5.7%, respectively. Notably, these enhancements are accomplished while maintaining the dimensions of the model and the parameter count. During the identification of defects on railroad track surfaces, our improved algorithm surpasses other widely used algorithms in terms of performance.

INDEX TERMS Rail defects detection, deep learning, YOLOv8, convolution module, attention mechanism, loss function.

I. INTRODUCTION

Due to the exponential increase in the railroad industry, there has been a constant expansion in operational mileage, speed, and density. Consequently, the safety risks associated with railroads are also on the rise [1]. This poses a greater challenge to the requirements of railroad inspection. The friction and rolling contact between high-speed trains and the rail surface can lead to wear, deformation, and, over time, the accumulation of corrugation, cracks, scars, rail breaks and other defects [2]. Corrugation refers to the periodic, sinusoidal wear or deformation observed on the surface of railway tracks. Cracks are characterized by linear or small fissures that manifest on the rail surface. Scars denote surface scratches or wear marks. Rail breaks represent fractures occurring at one or multiple points along the rail line. If these issues are not timely detected and repaired, they can jeopardize the safety of rail transportation, potentially resulting in train derailments and severe accidents. Hence, the prompt detection and timely repair of rail surface defects are of utmost importance. Doing so significantly reduces the risk of accidents, ensures transportation safety, extends the service life of rails, and reduces maintenance costs.

Traditional rail defect detection methods include manual detection [3], magnetic particle detection [4], and infrared thermography detection [5]. Manual inspection is a simple

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

and direct method, but it suffers from subjective judgment and fatigue of the inspector, leading to inconsistent and inaccurate results. Moreover, it is inefficient and costly. Magnetic particle inspection can visualize rail surface defects, which is beneficial for initial assessments. However, it involves complex operations, requires high technical expertise, and has strict environmental requirements. Similar to manual inspection, it is also prone to subjective assessments. Infrared thermography inspection, being a non-contact method, minimizes interference with the railroad track. However, it primarily focuses on heat-related defects and has limited capability to detect other types of defects, such as cracks. Additionally, it struggles to provide precise information regarding the size and depth of defects.

In the past few years, remarkable advancements have been achieved in the realm of artificial intelligence technology, with notable breakthroughs observed in the domain of machine vision. This progress has given rise to various neural network models that offer high accuracy and fast response times [6]. The introduction of these models presents a new solution for rail defect detection, allowing for significant reduction in human and material resource investments while enhancing the accuracy and efficiency of detection [7]. The model series known as You Only Look Once (YOLO) [8], as a widely utilized target detection framework, has been extensively employed in detecting railroad track defects, yielding commendable accuracy and detection outcomes. Alternatively, the Region-based Convolutional Neural Networks (Faster R-CNN) [9], another prevalent target detection model, leverages candidate region extraction and classification regression networks to precisely locate and identify defects on the track. Moreover, several investigations have amalgamated deep learning models with image segmentation techniques, enabling precise segmentation and detection of railroad track defects. Noteworthy examples encompass the adoption of models like U-Net [10] and Mask R-CNN [11] for localizing and segmenting defective regions on the track. Overall, deep learning algorithms [12] have exhibited substantial promise and yielded remarkable outcomes in the field of detecting defects in railroad tracks.

In the realm of railroad safety, the identification of surface abnormalities on rails assumes a pivotal role. In recent decades, machine vision technology has made remarkable progress, witnessing significant advancements in its application to the detection of railroad track defects across various countries. Internationally, Sresakoolchai and Kaewunruen [13] introduced a novel approach aimed at detecting track defects through the utilization of Track Geometry Correlation (TGC) to acquire precise geometric representations of the track. The proposed method leverages a Deep Neural Network (DNN) [14] model to effectively identify and classify defects. The experimental results showcase a remarkable accuracy rate of 92.17% achieved by the proposed method in detecting rail cross-section anomalies and worn-out rail spikes. However, the model's performance deteriorates as the number of defect categories increases. Mohan et al. [15] introduced an enhanced deep learning model, referred to as YOLOv2, which utilizes a dual-fold skip architecture for the recognition and defect detection of train bogie components in real-time video sequences. The model achieves an accuracy of 69.0%. The model's subpar performance can be attributed to the utilization of an older version of YOLO, which hinders its detection capabilities. Casas et al. [16] used the YOLOv8 [17] model to automatically detect and calculate eucalyptus stacked timber in forestry, using the CSPDarknet53 backbone network with a mAP50 of 83.9%. Nevertheless, it consistently exhibited a tendency to underestimate the quantity of stacked wood in the static images, resulting in errors ranging from -32.817% to -48.805%. In the domestic domain, Cao [18] developed a deep learning-based visual inspection system for surface defects on cold heavy rails. This system improves the parameters and structure of the Faster-RCNN target detection model. By adjusting the algorithmic logic, it effectively enhances defect detection efficiency and ensures the quality of heavy rail production. The proposed system exhibits a detection accuracy and recall rate exceeding 90% for both roll scar and roll mark test data. However, owing to the limited and imbalanced nature of the dataset employed by the system, the model encounters a notable overfitting phenomenon, thereby resulting in potential omissions and recognition errors during real-world testing scenarios. Bai et al. [19] proposed an enhanced YOLOv4-based method for detecting defects on railroad surfaces. The proposed approach employs MobileNetv3 as the underlying architecture for the extraction of image features within the YOLOv4 framework. Additionally, the PANet layer in YOLOv4 incorporates depth-separable convolution, leading to a notable accuracy improvement of 1.64% when compared to the original YOLOv4 model. An innovative approach for detecting track fastener defects was introduced by Wang et al [20], leveraging the YOLOV5 framework. To augment the model's capability, they employed the Filter Pruning via Geometric Median (FPGM) algorithm for model pruning, allowing for controlled increasing the model's width and depth. As a consequence, the mean average precision (mAP) witnessed a substantial increase from 91.23% to 93.42%. An enhanced YOLOX-Nano method for rail fastener defect detection was proposed by Hu et al. [21] by incorporating Adaptive Spatial Feature Fusion (ASFF) is applied immediately following the feature maps produced by the output of the PAFPN, and the mAP value of the improved YOLOX-Nano model increased by 18.75%. Wang et al. [22] introduced an improved road defect detection algorithm that integrates the BiFPN concept and reconstructs the neck structure of YOLOv8s within the framework, which improved the average accuracy by 3.3% in comparison with original model mAP@0.5.These advancements not just increase the accuracy of rail detection but also contribute to the modernization of China's railroad safety management system.

The detection of surface defects on railroad tracks presents several challenges that necessitate resolution [23]. Firstly, the presence of diverse and intricate background interferences, including rust, dirt, and coatings, poses a considerable difficulty as they bear resemblance to genuine defects, thereby exacerbating the detection process. Moreover, the manifold shapes, sizes, and textures exhibited by railroad track surface defects encompass a wide spectrum of variations, encompassing corrugations, abrasions, and fractures, thereby impeding the development of a generalized detection algorithm. Furthermore, the exigency for real-time or near real-time railroad track defect detection necessitates the implementation of detection algorithms characterized by high efficiency and rapid response capabilities. Lastly, practical implementations of the railroad track detection system are vulnerable to an array of external environmental disruptions, such as fluctuations in lighting conditions, as well as the presence of inclement weather conditions like rain and snow. These factors introduce perturbations to the image quality, consequently impinging upon the accuracy of defect identification.

YOLOv8 meticulously crafted by Ultralytics, represents a state-of-the-art (SOTA) model. Building on the accomplishments of prior iterations, YOLOv8 introduces novel features and enhancements, aiming to augment performance and versatility. Rooted in the principles of expeditiousness, precision, and user-friendliness, YOLOv8 emerges as an exemplary option for diverse applications encompassing target detection, image segmentation, and image classification tasks. In response to the challenge of detecting surface defects on railroad tracks, this study presents an improved algorithm that builds upon the foundation of YOLOv8. The algorithm presents the following notable innovations:

(a)In the backbone network of YOLOv8n, the conventional convolutional layer with the original stride of 2 is replaced with space-to-depth layer followed by a non-strided convolution layer (SPD-Conv) [24] building blocks. By implementing this modification, the model exhibits significantly improved sensitivity towards small and medium-sized targets.

(b)At each stage of downsampling, an efficient multiscale attention (EMA) module [25] is embedded after each C2f module. This module allows for the comprehensive utilization of feature information from different layers, the algorithm's feature characterization ability is significantly enhanced.

(c)The Focal-Segmentation Intersection over Union (Focal-SIoU) [26], [27] loss function is utilized as a replacement for the original Complete-Intersection over Union (C-IoU) loss function. This replacement adjusts the sample weights and increase the penalty for challenging samples that pose difficulties in classification. Consequently, the model's capability to accurately detect intricate samples is enhanced.

Through meticulous dataset annotation and rigorous validation processes, our algorithm achieves precise localization and accurate identification of defects on railroad track surfaces. These enhancements significantly improve the algorithm's overall detection performance.

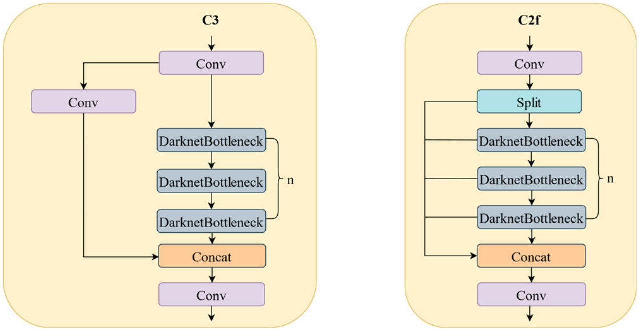

FIGURE 1. Schematic diagram of the C2f and C3 m.

II. IMPROVED ALGORITHM DESIGN

A. PRINCIPLES OF YOLOV8 ALGORITHM

The YOLOv8 algorithm, introduced by Glenn Jocher, builds upon and enhances the characteristics of the YOLOv5 algorithm. Additionally, this algorithm has developed an instance segmentation model using the You Only Look At Coefficients (YOLACT) [28] architecture. Similar to YOLOv5, YOLOv8 offers various model versions, encompassing a wide range of sizes, including nano, small, medium, large, and extra-large (n/s/m/l/x), depending on the requirements of different scales. To address the need for real-time inspection while considering the model's scale and parameter count, this study adopts the YOLOv8n defect detection model for identifying flaws on railroad track surfaces. The main improvements of YOLOv8 include:

(1) The kernel size used in the initial convolutional layer in YOLOv8 is modified from 6 × 6 to 3 × 3, while eliminating the two convolutionally connected layers in the neck module. As illustrated in Figure 1, all the C3 modules in the network are substituted with the new C2f module, which introduces additional branches to enhance the gradient flow and enrich the tributaries.

(2) The Head section has been considerably changed from YOLOv5 to the prevailing decoupled header structure in use [29], which segregates the classification and detection headers, as well as from Anchor-Based to Anchor-Free [30].

(3) For Loss computation, the TaskAlignedAssigner positive sample assignment strategy [31] was utilized. This strategy can be described as a matching approach that employs a weighted scoring mechanism to select positive samples. This approach incorporates both classification and regression scores to assign appropriate weights to the samples. Distribution Focal Loss was employed since the regression branch needed to be aligned with the integral form representation proposed in Distribution Focal Loss [32]. Additionally, CIoU Loss [33] was also incorporated.

(4) In the data augmentation step during training, an operation of disabling Mosiac enhancement [34] for the last

10 epochs in YOLOX was introduced, resulting in improved accuracy. Figure 2 illustrates the complete network architecture of YOLOv8.

B. IMPROVED METHOD BASED ON CONVOLUTIONAL MODULE SPD-CONV

To bolster the capability to extract features from the YOLOv8n model, this paper incorporates the SPD-Conv convolution module within the backbone network. When tackling low-resolution images and compact-sized targets, the conventional YOLOv8 model may face certain limitations that can lead to a decline in its detection performance. Since the railroad track surface defects often involve a significant number of small targets drop defects, the original YOLOv8 model is ineffective in detecting such targets. The incorporation of the SPD-Conv convolution module in the YOLOv8 model leads to a substantial improvement in feature representation, all while preserving the model's overall structure, thereby reducing the reliance on inputs of good quality.

The SPD-Conv module serves as a complete replacement for step convolution and pooling layers. It comprises a space-to-depth (SPD) layer followed by a non-stepwise convolution (Conv) layer. More specifically, the input feature maps are initially transformed using the SPD layer and subsequently undergoes convolution operation through the non-stepwise convolution layer. This combination effectively reduces the spatial dimension without sacrificing information, while preserving channel information. Consequently, the integration of the SPD-Conv module into the CNN yields substantial improvements in the detection performance, particularly when the network is faced with challenges presented by low-resolution images and compact-sized targets.

To obtain sub-feature mappings from a mid-level feature representation of size $S \times S \times C_1$, the following sequence slicing procedure is employed:

$$\begin{aligned}
f_{0,0} &= X\left[0:S:scale, 0:S:scale\right], \\
f_{1,0} &= X\left[1:S:scale, 0:S:scale\right], \cdots, \\
f_{scale-1,0} &= X\left[scale-1:S:scale, 0:S:scale\right]; \\
f_{0,1} &= X\left[0:S:scale, 1:S:scale\right], \\
f_{1,1} &= X\left[1:S:scale, 1:S:scale\right], \cdots, \\
f_{scale-1,1} &= X\left[scale-1:S:scale, 1:S:scale\right]; \\
&\vdots \\
f_{0,scale-1} &= X\left[0:S:scale, scale-1:S:scale\right], \\
&f_{1,scale-1}, \cdots, \\
f_{scale-1,scale-1} &= X\left[scale-1:S:scale, scale-1:S:scale\right]
\end{aligned} \tag{1}$$

We define a subgraph $f_{x,y}$ to consist of all $X(i, j)$ that satisfy both $i + x$ and $j + y$ being integral divisors of the *scale*. Thus, the *scale* of downsampling is applied to each subgraph. Example results are illustrated in Figure 3(a)(b)(c) to demonstrate the effect of setting the $scale = 2$. By downsampling X by scale factor 2, we obtain four subgraphs $f_{0,0}, f_{1,0}, f_{0,1}, f_{1,1}$. The shape of each subgraph is $(S/2, S/2, C_1)$.

Following that, a new feature representation called X' is created by connecting these sub-feature representations across the channel dimension. This new feature map has been processed by scaling down the spatial dimension with a reduced scale factor, while in the dimension of the channel it has been processed by a scale squared factor. In Figure 3(d), it is evident that the SPD operation transforms the original feature representation $X(S, S, C_1)$ into a mid-level feature representation denoted as $X'\left(S/scale, S/scale, scale^2 C_1\right)$.

Following the SPD feature conversion layer, we incorporate a non-spanning convolutional layer, specifically a step-size 1 convolutional layer. This convolutional layer is equipped with C_2 filters, where $C_2 < scale^2 C_1$, which continues the conversion from $X'\left(S/scale, S/scale, scale^2 C_1\right)$ to $X''(S/scale, S/scale, C_2)$ and the transformation process is illustrated in Figure 1.3(e). By employing a non-spanning convolution, the objective is to retain a significant amount of information pertaining to all discriminative features. The utilization of a 3 × 3 filter with a stride of 3 leads to a scenario where each pixel within the feature map is sampled exactly once. This process produces a distinct "shrinking" effect on the feature map. Likewise, employing a stride of 2 introduces an asymmetrical sampling pattern, where the even and odd columns are treated differently during the data extraction process. It is crucial to highlight that when using a stride larger than 1, there is a tendency for the model to experience a loss of discriminative information.

C. EMA ATTENTION MECHANISM

In railroad track surface defect detection, the traditional YOLOv8n model's simple feature fusion strategy faces challenges when dealing with coexisting large-scale and small-scale targets. This strategy tends to limit the depth of feature representation. While the significance of the attention mechanism in enhancing feature representation is widely recognized, traditional channel dimensionality reduction methods may compromise the integrity of deep visual information. The EMA attention mechanism, which avoids dimensionality reduction, achieves comprehensive information retention and computational efficiency by reconstructing a portion of the channels and uniformly distributing spatial semantics among sub-features. It not only globally encodes information to adjust channel weights but also captures pixel-level relationships through cross-dimensional interactions. Within this study, the EMA module is integrated into the neck section of YOLOv8n to effectively address the challenge of multi-scale target detection. This integration significantly improves the model's performance in detecting railroad track surface defects.

The Coordinate Attention (CA) attention module can be viewed as a comparable alternative to the SE attention module, as both mechanisms aim to incorporate cross-channel information by utilizing global average pooling operations. In general, the utilization of global average pooling serves the

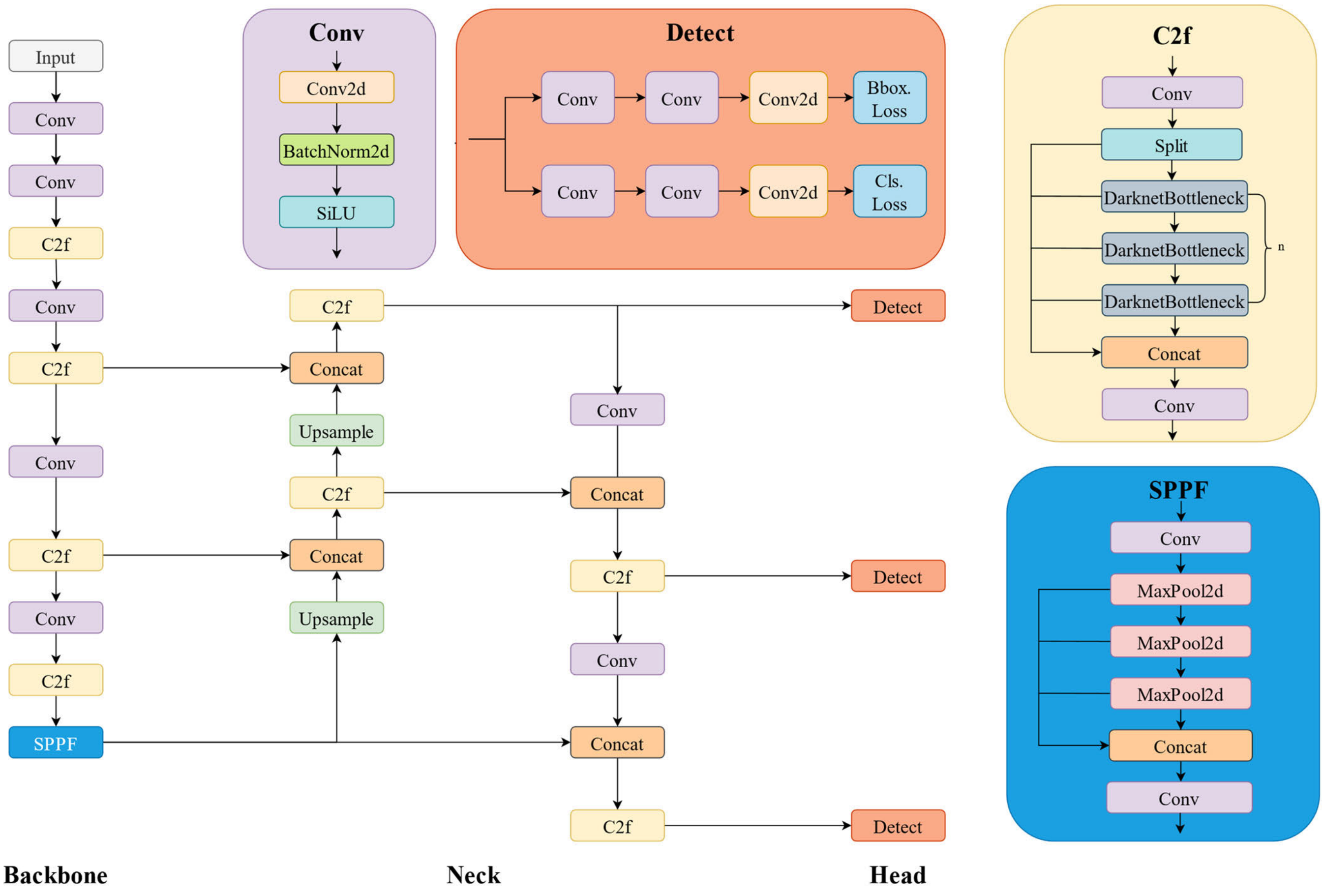

FIGURE 2. YOLOv8 network architecture.

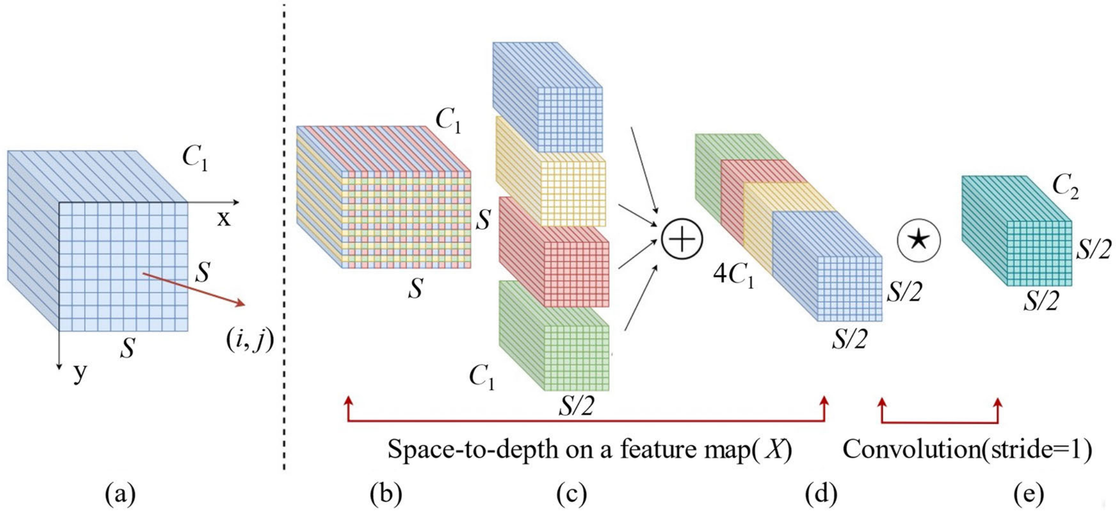

FIGURE 3. SPD-Conv with graphically represented.

purpose of generating channel-wise statistics by compressing global spatial location information into channel descriptors.

Slightly different from Squeeze-and-Excitation (SE), the CA attention module integrates spatial location information into

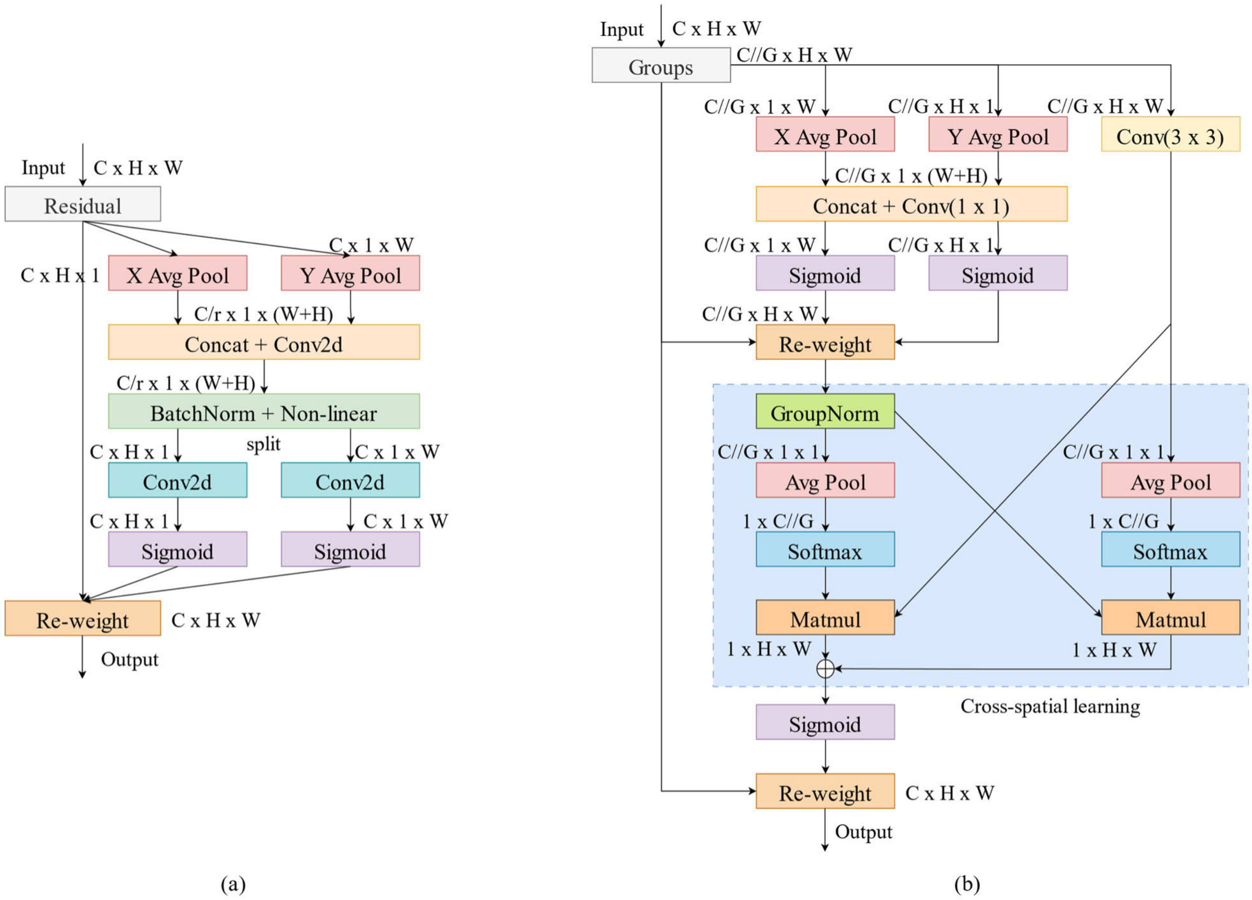

FIGURE 4. CA and EMA network structure.

the channel attention graph to improve the consolidation of features. Figure 4(a) illustrates the architecture of the CA attention module.

Figure 4(b) provides a representation of the comprehensive structure of the EMA attention module. While the CA attention module achieves good performance by incorporating spatial information into channel modeling, it overlooks the interaction between complete spatial locations. Additionally, the restricted receptive field of the 1 × 1 convolution impairs the ability to interact locally across channels and utilize contextual information. Instead, the EMA module selects the mutual elements of the 1 × 1 convolution within the CA module and refers to them as the 1 × 1 branches. In order to effectively consolidate spatial structure information across multiple scales, the introduction of a 3 × 3 kernel, which operates in parallel with the 1 × 1 branch, facilitates rapid response and is referred to as the 3 × 3 branch. By employing a parallel sub-network structure, the EMA module effectively preserves precise spatial structural information within each channel. Simultaneously, it captures inter-channel information to regulate the significance of individual channels. Additionally, the EMA employs a method of aggregating interspatial information using various directions across spatial dimensions, which effectively enhances the aggregation of features. To achieve this, the procedure includes the introduction of two tensors: one for the output of the branch with a 1 × 1 dimension and another for the output of the branch with a 3 × 3 dimension. In order to capture overall spatial knowledge within the exports of the 1 × 1 branch, a 2D global mean pooling operation is employed. Prior to the collaborative activation mechanism that incorporates the channel characteristics, the output of the smallest branch undergoes a direct reshaping process to conform to the appropriate dimensional structure. To denote the operation of 2D global pooling, the following representation is used:

$$z_c = \frac{1}{H \times W} \sum_{j}^{H} \sum_{i}^{W} x_c(i, j) \tag{2}$$

For efficient computation, the exports of the 2D global average pooling undergo a nonlinear transformation using the softmax function. To ensure efficiency and compatibility with modern architectures, the exports obtained from the EMA module are specifically designed to match the size of X.

D. FOCAL_SIOU LOSS FUNCTION

To tackle the issue of imbalanced positive and non-positive instances in single-stage target detection, the Focal Loss is employed as the chosen loss function. By adjusting the assigned weights for positive and negative samples, the Focal

Loss places a higher emphasis on samples that are challenging to classify. By employing the Focal Loss is selected as the designated loss function, the model's capacity to identify complex instances is enhanced. SIoU enhances the stability and accuracy of detecting small and partially occluded targets by smoothing the IoU metric. In the context of railroad track surface defect detection, various types of targets may overlap or occlude each other, and there is a significant number of small target defects. Therefore, both Focal Loss and SIoU, as two improved techniques, are simultaneously applied to the YOLOv8 model. By enabling the network to allocate increased attention to individual target instances, the model's performance and effectiveness are enhanced.

To address the challenge of imbalanced class distributions, Focal Loss incorporates a modulation factor that adjusts the weight assigned to each sample. The modulation factor is determined using the following formula:

$$FL\left(p_t\right) = -\left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \tag{3}$$

where γ is a parameter within the range of [5, 0], and the modulation factor $(1 - p_t)^{\gamma}$ reduces the emphasis on easily categorizable samples in the loss function. Focal Loss enhances the weight assigned to difficult-to-classify instances, favoring their contribution to the loss. This helps improve the accuracy of difficult-to-classify instances. When p_t is larger (indicating better assignment of easier samples), the paired loss becomes smaller. By achieving an equilibrium between positive and non-positive instances, as well as balancing the difficulty levels of instances, the final formulation of the Focal Loss can be obtained:

$$FL\left(p_t\right) = -\alpha_t \left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \tag{4}$$

Parameter α_t can be utilized to address the disproportionate distribution between the quantity of positive and non-positive instances, providing a means to mitigate this disparity. On the other hand, parameter γ can be employed to control the imbalance between quantity of easily categorizable instances and the quantity of difficult-to-categorize samples. Determining the optimal weight ratio typically necessitates empirical exploration and fine-tuning. Through systematic experimentation of various weight ratios and evaluating model performance on validation sets or through cross-validation, the weight ratio that achieves a more favorable equilibrium is chosen.

Loss functions like Generalized-Intersection over Union(GIoU), Distance-Intersection over Union(DIoU), CIoU, and others do not consider the spatial alignment between the ground truth bounding box and the predicted bounding box, leading to slower convergence speed. To tackle this concern, the SIoU loss function accounts for consideration the angle formed in terms of spatial alignment between the ground truth bounding box and the predicted bounding box, introducing a novel approach to redefine the related loss function. The redefined loss function comprises the following components:

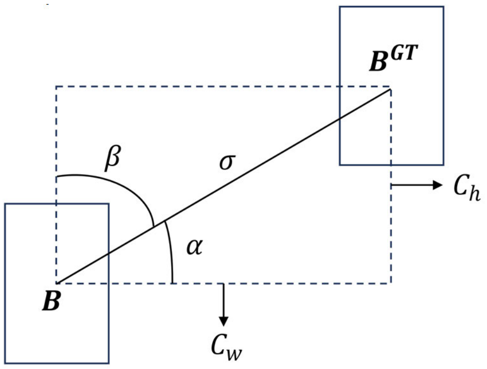

FIGURE 5. SIoU schematic diagram.

(1) The angle cost represents the minimal angle established by the line connecting the centroid of the objects and the axis of reference. It is defined as follows:

$$\begin{aligned}\Lambda &= 1 - 2 \times \sin^2\left(arc\sin\left(\frac{c_h}{\sigma}\right) - \frac{\pi}{4}\right) \\ &= \cos\left(2 \times \left(arc\sin\left(\frac{c_h}{\sigma}\right) - \frac{\pi}{4}\right)\right)\end{aligned} \tag{5}$$

As depicted in Figure 5, the height difference denoted by c_h refers to the disparity in height observed between the centroids of the ground truth bounding box and the predicted bounding box. Additionally, σ indicates the distance between the centroids of the ground truth and predicted bounding boxes. The following definitions are assigned to the respective values:

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2} \tag{6}$$

$$c_h = max\left(b_{c_y}^{gt}, b_{c_y}\right) - min\left(b_{c_y}^{gt}, b_{c_y}\right) \tag{7}$$

In this discussion, the coordinate notation is used to represent the center position of the ground truth bounding box $\left(b_{c_x}^{gt}, b_{c_y}^{gt}\right)$, the coordinates $\left(b_{c_x}, b_{c_y}\right)$ represent the center position of the predicted bounding box. The angle loss is 0 when α is $\pi/2$ or 0.

(2) The distance cost refers to the measurement of the spatial separation between the centroids of the bounding boxes and its penalty cost is positively correlated with the angle cost. It is defined as follows:

$$\Delta = \sum_{i=x,y}\left(1 - e^{-\gamma\rho_t}\right) = 2 - e^{-\gamma\rho_x} - e^{-\gamma\rho_y} \tag{8}$$

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda \tag{9}$$

where (c_w, c_h) represents the dimensions of the smallest enclosing rectangle, comprising the width and height of both the ground truth and predicted bounding boxes.

(3) The shape cost is calculated by computing the disparity in width between the two boxes, as well as the relative proportion of their maximum width (and length). This enables the overall shape convergence to align with the convergence of

the longer and wider sides. The following definition outlines the concept:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-w_t}\right)^{\theta} = \left(1 - e^{-w_w}\right)^{\theta} + \left(1 - e^{-w_h}\right)^{\theta} \tag{10}$$

$$w_w = \frac{\left|w - w^{gt}\right|}{max\,(w, w^{gt})},\ w_h = \frac{\left|h - h^{gt}\right|}{max\,(h, h^{gt})} \tag{11}$$

By measuring the difference in width between the predicted bounding box and the ground truth bounding box, their width and height are represented by the coordinate pairs (w, h) and $\left(w^{gt}, h^{gt}\right)$, respectively. The parameter θ controls the level of emphasis on the shape loss. To alleviate the excessive emphasis on shape loss and minimize the displacement of the predicted bounding box, parameter θ is constrained within the range of [2] and [6].

(4) The IoU loss is defined as follows [35], encompassing:

$$IoU = \frac{B \bigcap B^{GT}}{B \bigcup B^{GT}} \tag{12}$$

The following expression outlines the formulation of the final SIoU loss function:

$$Loss_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{13}$$

E. NETWORK STRUCTURE

To tackle the common issue of frequent dropout of small targets in rail surface defect detection, this paper presents a proposed improved network architecture. The architecture, depicted in Figure 6, is specifically designed to tackle this issue effectively. Due to the unsatisfactory performance of the standard YOLOv8 model in recognizing small targets, we propose the SPD-Conv convolution module. By strengthening the model's capacity to extract pertinent features and reducing reliance on high-quality input data, notable improvements are achieved. Specifically, within the YOLOv8 backbone network, we substitute the original convolutional layer with a stride of 2 with a block constructed using SPD-Conv. This modification allows for more efficient capture of details in small and medium scale targets. Additionally, considering the presence of targets at different scales in rail surface defect detection, we introduce the EMA attention block after each C2f block in the downsampling stage of the model. This ensures effective utilization of feature information at various detection levels, thereby improving the accuracy of feature representation. Furthermore, the Focal-SIoU loss function is utilized as a substitute for the C-IOU loss function used in YOLOv8. This enhancement enables the model to better recognize challenging samples by adjusting the weighting of positive and negative samples and applying higher penalties to indistinguishable samples. Additionally, it encourages the model to focus more on individual target instances, further improving the accuracy of detection. The proposed approach in this research offers strengthened accuracy without compromising the model size or the number of parameters. This characteristic makes it particularly well-suited for the task of railroad track surface defect detection.

III. EXPERIMENT AND RESULT

A. DATASETS

In this research, a new dataset for railroad track surface defect detection was created by carefully selecting and preprocessing track defect detection [36] and clipping datasets [37] from open sources. The datasets pertaining to track defect detection and clipping were procured from the publicly accessible dataset repository, Roboflow Universe. The dataset associated with track defect detection was contributed by the user SUSTECH, while the dataset related to clipping was submitted by the user FENG. The original track defect detection dataset consisted of 3041 images, with 2924 training images and 117 validation images. The clipping dataset provided 4110 images, out of which 3963 were used for training and 294 for validation. However, due to issues such as duplicate images and an abundance of defect-free railroad track images in these open-source datasets, a rigorous screening and meticulous re-labeling process was conducted. As a result, a total of 3,812 high-quality images were selected to create the dedicated railroad track surface defect detection dataset for this study.

The dataset comprises five distinct categories for detection, namely cracks, scars, breaks, lightbands, and rails. Among these categories, as shown in figure 7(a), cracks manifest as linear or fine fractures appearing on the surface of the rail. The presence of cracks can induce deformations in the surrounding rail surface, such as minute elevations or depressions, resulting from stress accumulation and deformations caused by crack formation and expansion. As shown in figure 7(b), breaks correspond to cracks that emerge at single or multiple points on the rail surface. The cross-section of a broken rail may exhibit diverse characteristics, such as a flat, rough, or visibly cracked profile. As shown in figure 7 (c), scars are characterized by scraping or abrasion marks on the rail surface, which can assume various forms, including linear, punctual, or extensive areas. As shown in figure 7 (d), lightbands denote relatively bright areas that manifest as bands along the length of the rail and are typically attributed to friction, abrasion, or heat generation resulting from wheel-rail contact. Although lightbands generally do not pose an immediate safety threat to rail transportation, they can serve as early indicators of other, more severe defects. Therefore, timely detection and monitoring of lightbands can facilitate preventive maintenance and mitigate further wear and tear, ensuring the rail's optimal condition and smooth train operation. As shown in figure 7 (e), Rails are typically elongated strips characterized by specific lengths and shapes. In this study, in addition to detecting four defect types, namely cracks, scars, breaks, and lightbands, the detection and localization of rails are also performed. To ensure comprehensive training and evaluating the model, the dataset was partitioned

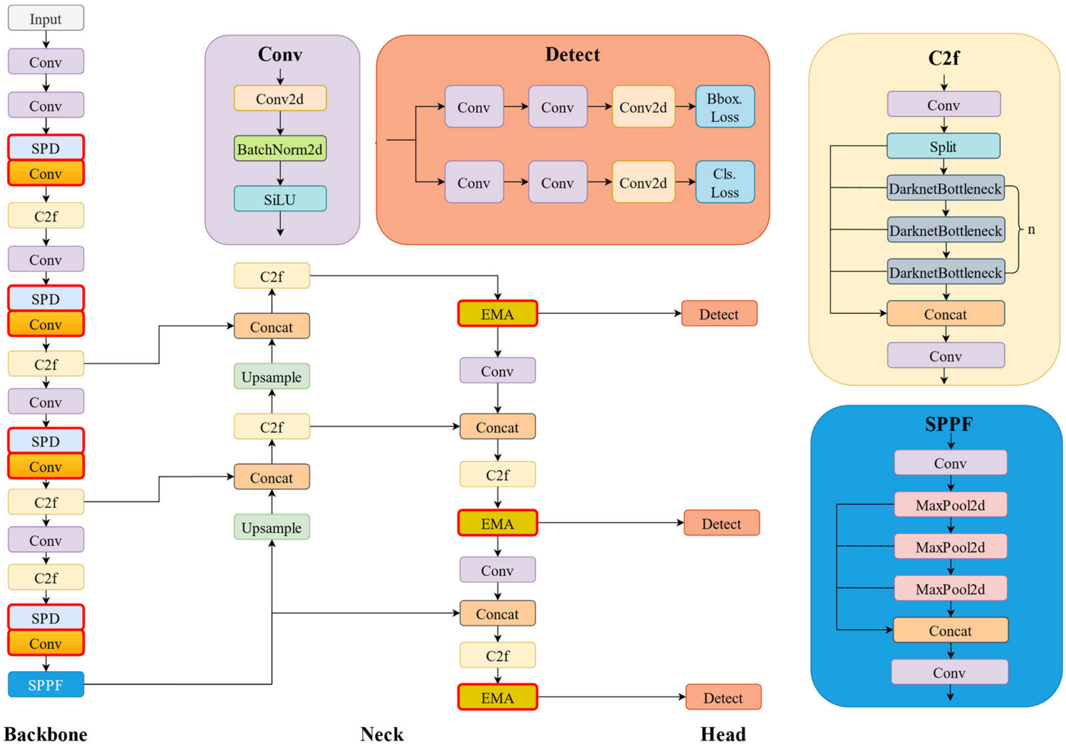

FIGURE 6. Improvements to the YOLOV8 network structure.

TABLE 1. Sample composition of the rail surface defect detection dataset.

Sample category	Number of training sets	Number of validation sets	Number of test sets	Total
cracks	3006	350	392	3748
breaks	455	47	61	563
scars	3313	421	384	4118
lightbands	3789	469	470	4728
rails	3831	483	457	4771

into training, validation, and test sets in an 8:1:1 ratio. For more detailed information on the dataset composition, please refer to Table 1.

B. EXPERIMENTAL SETUP

The performed experiments conducted in this investigation utilized the PyTorch framework and Graphics Processing Unit (GPUs). Table 2 provides a comprehensive breakdown of the configuration of the experimental conditions.

C. PERFORMANCE INDICATORS

When evaluating target detection algorithms, the primary assessment criteria can be categorized into two main groups: detection precision and model complexity. To evaluate the precision of the detection, metrics such as precision (P), recall (R), and mean average precision (mAP) are widely used.

Taking TP to represent the amount of true positive samples, FP to represent the amount of false positive samples, and FN to represent the amount of false negative samples. The mathematical expressions for precision (P), recall (R), and mean average precision (mAP) can be obtained through the following derivations:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$AP = \int_0^1 P(R)\, dR \tag{16}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \tag{17}$$

The algorithm's model complexity is determined by factors such as the model's dimensions, the number of parameters, and the computational requirements. Larger values in these factors indicate higher model complexity.

D. EXPERIMENTAL RESULTS AND ANALYSIS

1) TRAINING CURVES

To provide a more intuitive visualization of the enhancements achieved by the improved algorithm, we present the training curves. Figure 8 illustrates the mAP50, training loss, and

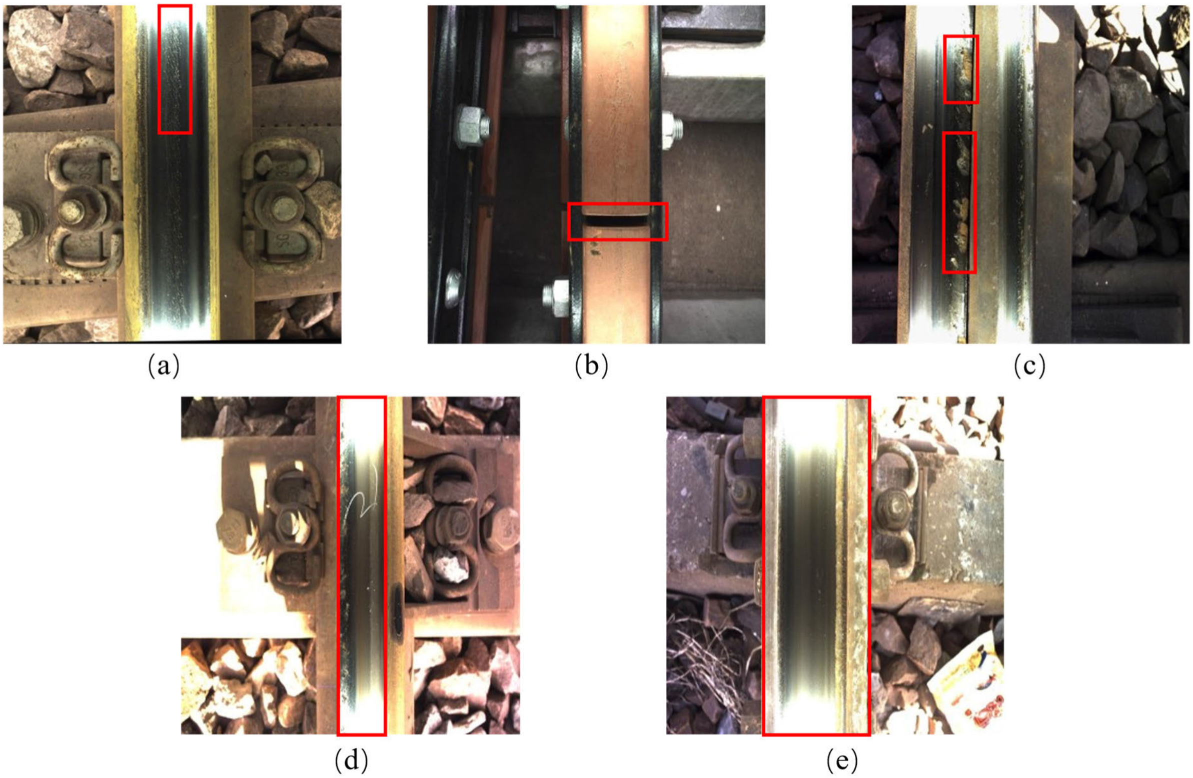

FIGURE 7. Schematic diagram of the 5 detecting categories.

TABLE 2. Configuration of the experimental conditions.

Name	Specific information
Operating system	Linux(Ubuntu)
CPU	Intel(R) Xeon(R) Platinum 8163 CPU @2.50GHz
GPU	TITAN Xp
RAM	12GB
CUDA	12.0
PyTorch	1.11.0
Python	3.8.16

validation loss curves obtained after 500 rounds of training for both the original and improved models. Notably, the training process for the original model concludes at 450 rounds since its accuracy ceases to improve further. In contrast, the improved model demonstrates accelerated convergence, resulting in predictions that align more closely with the ground truth values. Moreover, the mAP50 metric exhibits a substantial improvement, as clearly indicated by the aforementioned curves. This observed enhancement serves as convincing proof of the effectiveness of the algorithm being proposed.

2) ABLATION EXPERIMENT

To gauge the performance enhancement achieved by incorporating three optimization strategies into YOLOv8n, namely the SPD-Conv module, EMA module, and Focal-SIoU loss function, a sequence of ablation experiments is conducted on the dataset in this research work. Evaluation metrics such as model size, number of parameters, precision rate (P), recall rate (R), average precision rate (mAP@0.5), and average accuracy (mAP@0.5:0.95) are utilized. The ablation experiments involve different combinations of the aforementioned improvement modules. Table 3 showcases the results obtained from the experiments conducted in this study.

According to Table 3, it is plausible to observe that after integrating the SPD-Conv module into the original YOLOv8n model, there is a boost of 1.3%, 3.3%, 2.9%, and 2.4% in P, R, mAP@0.5, and mAP@0.5:0.95, respectively. However, after incorporating the CA mechanism, P experiences a decrease of 0.4% and mAP@0.5:0.95 witnesses a decrease of 0.5%. On the other hand, R exhibits an increase of 0.2%, and mAP@0.5 shows a slight improvement of 0.1%. Furthermore, after adding the EMA attention mechanism, there is a boost of 0.7%, 2.8%, 2.3%, and 2.5% in P, R, mAP@0.5, and mAP@0.5:0.95, respectively. Based on the aforementioned findings, it can be deduced that the incorporation of the EMA module yields superior outcomes compared to the integration of the CA module. Specifically, this augmentation leads to notable enhancements across various evaluation metrics, with a boost of 1.1%, 2.6%, 2.2%, and 3.0% in P, R, mAP@0.5, and mAP@0.5:0.95, respectively. Additionally, after using the Focal-SIoU loss function, there is a boost of 3.4%, 2.1%, 2.7%, and 2.4% in P, R, mAP@0.5, and mAP@0.5:0.95, respectively. The experimental results showcased in this table

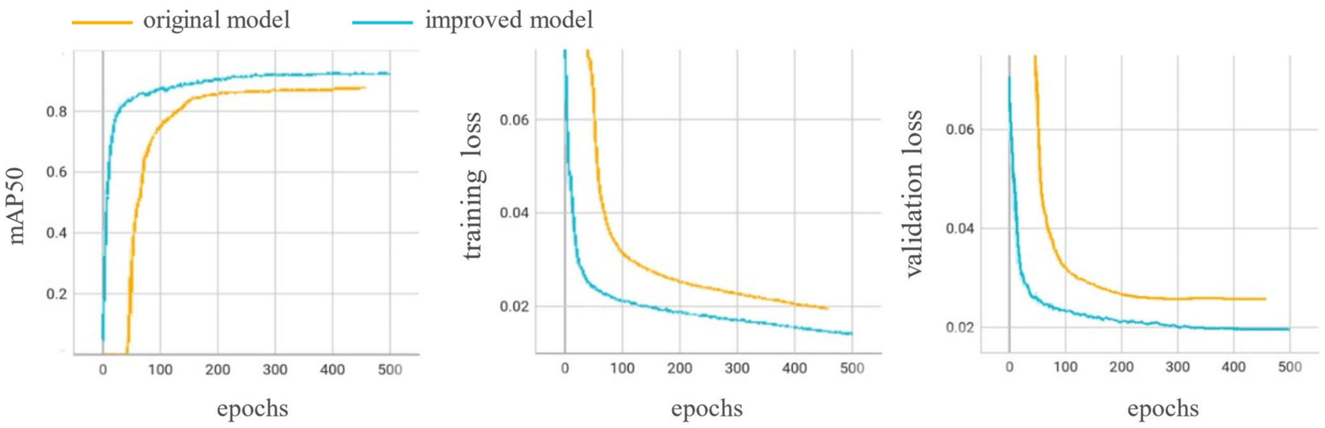

FIGURE 8. Comparison of training curves before and after model improvement.

TABLE 3. Ablation experiment.

Methods					Model size/MB	Params/ M	P /%	R /%	mAP@0.5 /%	mAP@0.5:0.95 /%
YOLOv8n	SPD-Conv	EMA	CA	Focal-SIoU						
√					**6.2**	3.0	87.5	86.4	86.3	71.1
√	√				6.8	3.3	88.8	89.7	89.2	73.5
√		√			6.3	3.0	88.2	89.2	88.6	73.6
√			√		6.3	3.0	87.1	86.6	86.4	70.6
√				√	6.8	3.0	90.9	88.5	89	73.5
√	√	√			6.8	3.3	91.1	90.2	90.8	72.3
√	√			√	6.8	3.3	90.9	89.5	90.2	74.7
√		√		√	6.3	3.0	89.5	89.6	89.3	73.8
√	√	√		√	6.8	3.3	**93.1**	**91.3**	**91.5**	72.8

TABLE 4. Comparison experiments with other algorithms.

Algorithms	Model size/MB	Params /M	P /%	R /%	mAP@0.5 /%	mAP@.5:.95 /%
SSD	96.7	24.0	90.2	45.7	74.0	48.9
YOLOv5n	3.9	1.8	89.9	82.5	86.8	65.7
YOLOv6n	_	4.6	89.5	78.4	89.4	72.4
YOLOv7-tiny	12.3	6.0	88.1	84.1	87.7	67.7
YOLOv8n	6.2	3.0	90.3	88.7	88.4	73.3
Proposed Algorithm	**6.8**	**3.3**	**93.9**	**93.7**	**94.1**	**73.5**

provide evidence of the improved model proposed in this paper, with a slight increase in model size and parameters. P, R, mAP@0.5, and mAP@0.5:0.95 are boosted by 5.6%, 4.9%, 5.2%, and 1.7%, respectively. This serves as evidence of the efficacy of the algorithm enhancement proposed in this research study.

3) COMPARISON EXPERIMENT

To substantiate the superiority of the enhanced algorithm proposed in this research work, a comparative experimental comparison was performed under identical conditions, involving several widely-used target detection algorithms. The evaluation metrics employed were model size, parameters, precision (P), recall (R), average precision (mAP@0.5), and average precision (mAP@0.5:0.95). The experimental results are presented in Table 4, which clearly illustrates that the algorithm suggested in this research work achieves enhanced accuracy in target detection while maintaining the intricacy of the model. By comparing the outcomes of the enhanced algorithm with those of the original YOLOv8n model, it becomes apparent that the proposed approach exhibits substantial advancement, as evidenced by a notable increase of 5.7% in mAP@0.5. This substantial enhancement across all evaluation indicators further solidifies the algorithm's superiority over the existing approaches.

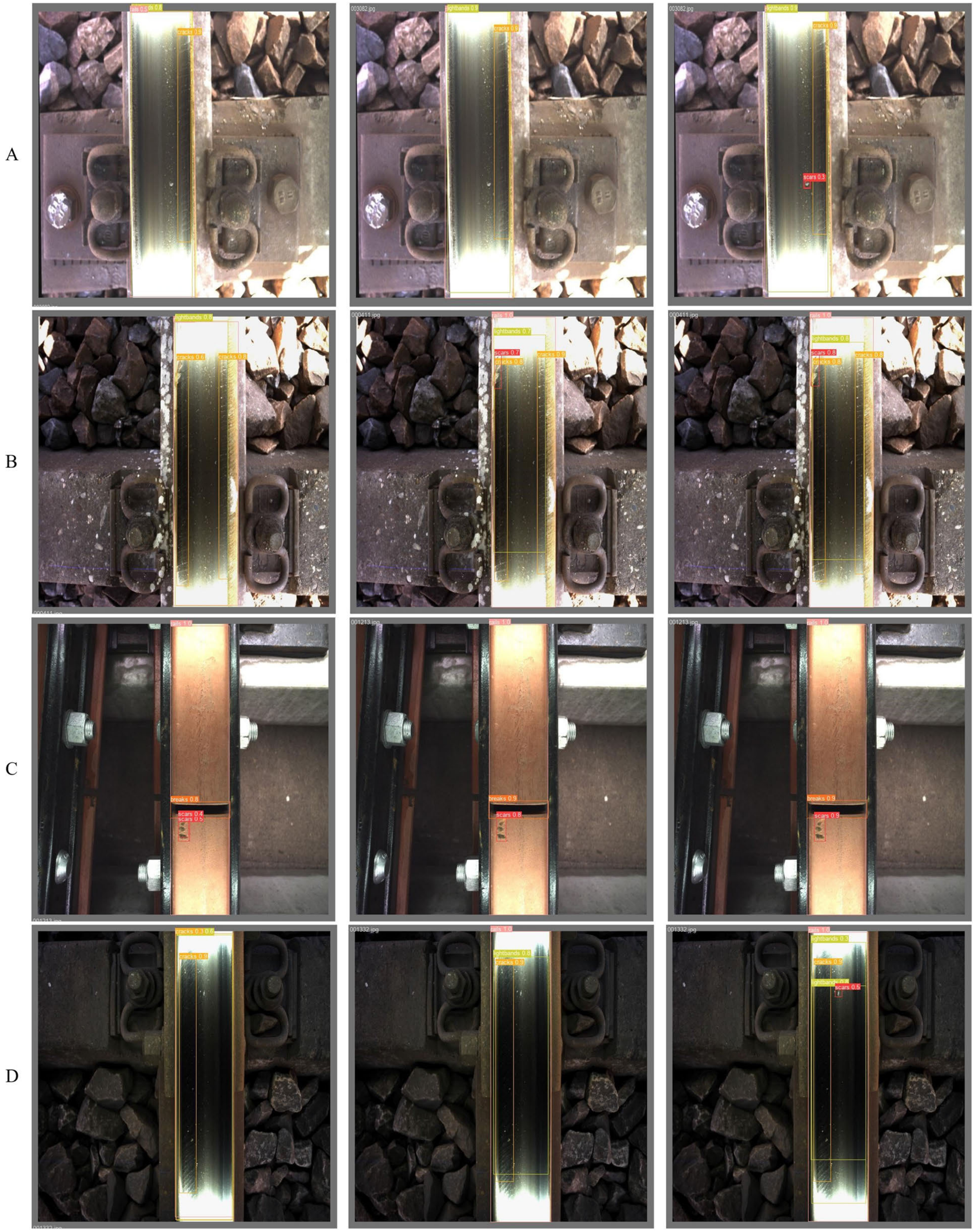

FIGURE 9. Comparison of rail surface defect detection algorithms.

4) ALGORITHM VALIDATION

Figure 9 showcases a comparative analysis of defect detection results on railroad track surfaces using visualized images obtained from YOLOv5n, YOLOv8n, and the improved YOLOv8n algorithms. The images encompass various lighting conditions, including those within tunnels. Through experimental comparisons within groups A and D, it is evident that the optimized YOLOv8n algorithm outperforms other models by detecting targets that remain unidentified by alternative approaches. Furthermore, the experimental comparison between Group B and Group C demonstrates the improved YOLOv8n algorithm's superior detection accuracy. The results indicate that the improved model achieves better detection outcomes across diverse environments, showcasing its generalization ability and robustness. These observations substantiate the algorithm's efficacy in effectively addressing challenges associated with detecting small and densely occluded targets, providing empirical evidence of its performance.

IV. CONCLUSION

This study introduces a novel detection algorithm built upon YOLOv8n, aiming to effectively tackle the challenge of detecting defects on the surface of railroad tracks. The algorithm incorporates several key enhancements to enhance its defect detection capabilities. In this algorithm, the standard convolutional layer with a step size of 2 in the backbone network is replaced with the SPD-Conv building block. The specific aim of this substitution is to enhance the detection performance for targets with small to medium dimensions. Moreover, to further enhance the algorithm's efficiency in utilizing feature information from different layers and its feature expression capability, the EMA attention module is introduced after each C2f module at each downsampling stage. Additionally, the original C-IOU loss function of YOLOv8 is substituted with the Focal-SIoU loss function. By adjusting the weights assigned to samples representing positive and negative categories, the penalty imposed on challenging samples is amplified. Thus, the model's recognition accuracy for complex samples is upgraded, consequently improving the overall performance of the algorithm. Through attaining accurate pinpointing and detection of flaws on the surface of the railroad track by means of meticulous dataset screening, labeling, and validation procedures. The detection performance of the algorithm experiences a significant enhancement due to these improvements.

The results obtained from the improved algorithm are remarkably impressive, with precision, recall, and average accuracy achieving 93.9%, 93.7%, and 94.1%, respectively. These values exhibit substantial improvements of 3.6%, 5.0%, and 5.7% over the original YOLOv8n model. It is worth emphasizing that these advancements were accomplished without augmenting the model size or parameters, underscoring the algorithm's efficacy and utility in detecting surface defects on railroads. To further substantiate the algorithm's detection capabilities, simulation experiments were conducted in diverse detection environments to assess the model's generalization capability and robustness. The detection of railroad tracks, encompassing various lighting conditions and tunnels, was performed. The outcomes consistently demonstrate the superior detection performance of the improved algorithm, surpassing that of other mainstream detection algorithms within this domain. Future efforts will be dedicated to refining the network architecture to enhance detection accuracy and processing speed. Additionally, plans are underway to adapt the model for deployment on edge computing platforms, necessitating the porting and optimization of the algorithm to reduce its size and facilitate seamless deployment.

REFERENCES

[1] Y. Zhao, Z. Liu, D. Yi, X. Yu, X. Sha, L. Li, H. Sun, Z. Zhan, and W. J. Li, "A review on rail defect detection systems based on wireless sensors," *Sensors*, vol. 22, no. 17, p. 6409, Aug. 2022.

[2] J. H. Feng, H. Yuan, Y. Q. Hu, J. Lin, S. W. Liu, and X. Luo, "Research on deep learning method for rail surface defect detection," *IET Electr. Syst. Transp.*, vol. 10, no. 4, pp. 436–442, Dec. 2020.

[3] J. W. Huo, Z. Liu, and Y. D. Wang, "Planar electromagnetic tomography for rail flaw detection," *Chin. J. Electr. Eng.*, vol. 41, no. 15, pp. 5351–5361, 2021.

[4] X. Q. Hu, "Research on the application of nondestructive testing technology in steel structure construction engineering," *Real Estate World*, vol. 10, no. 19, pp. 142–144, Oct. 2023.

[5] K. Tomita and M. Y. L. Chew, "A review of infrared thermography for delamination detection on infrastructures and buildings," *Sensors*, vol. 22, no. 2, p. 423, Jan. 2022.

[6] S. Orlov, A. Piletskaya, N. Kusakina, and A. Tyugashev, "Machine learning of diagnostic neural network for railway track monitoring," in *Cyber-Physical Systems: Intelligent Models and Algorithms* (Studies in Systems, Decision and Control), vol. 417, A. G. Kravets, A. A. Bolshakov, and M. Shcherbakov, Eds. Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-030-95116-0_5.

[7] Z. Chen, Q. Wang, Q. He, T. Yu, M. Zhang, and P. Wang, "CUFuse: Camera and ultrasound data fusion for rail defect detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21971–21983, Nov. 2022.

[8] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.

[9] X. Xu, M. Zhao, P. Shi, R. Ren, X. He, X. Wei, and H. Yang, "Crack detection and comparison study based on faster R-CNN and mask R-CNN," *Sensors*, vol. 22, no. 3, p. 1215, Feb. 2022.

[10] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[11] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: A survey," in *Computational Intelligence in Pattern Recognition*. Singapore: Springer, 2020, pp. 657–668.

[12] Z. Tu, S. Wu, G. Kang, and J. Lin, "Real-time defect detection of track components: Considering class imbalance and subtle difference between classes," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[13] J. Sresakoolchai and S. Kaewunruen, "Railway defect detection based on track geometry using supervised and unsupervised machine learning," *Struct. Health Monitor.*, vol. 21, no. 4, pp. 1757–1767, Jul. 2022.

[14] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: A review and experimental comparison," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4403–4462, Aug. 2022.

[15] K. K. Mohan, C. R. Prasad, and P. V. V. Kishore, "YOLO v2 with bifold skip: A deep learning model for video based real time train bogie part identification and defect detection," *J. Eng. Sci. Technol.*, vol. 16, no. 3, pp. 2166–2190, Jun. 2021.

[16] G. G. Casas, Z. H. Ismail, M. M. C. Limeira, A. A. L. da Silva, and H. G. Leite, "Automatic detection and counting of stacked eucalypt timber using the YOLOv8 model," *Forests*, vol. 14, no. 12, p. 2369, Dec. 2023.

[17] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with YOLOv8," 2023, *arXiv:2305.09972*.

[18] H. P. Cao, "Development of visual inspection system for surface defects of cold heavy rail based on deep learning," Zhejiang Univ., Hangzhou, China, Tech. Rep. Issue 06, 2021, doi: 10.27461/d.cnki.gzjdx.2020.003816.

[19] T. Bai, J. Gao, J. Yang, and D. Yao, "A study on railway surface defects detection based on machine vision," *Entropy*, vol. 23, no. 11, p. 1437, Oct. 2021.

[20] X. Wang, J. Zhang, Y. Wang, M. Li, and D. Liu, "Defect detection of track fasteners based on pruned YOLO v5 model," in *Proc. IEEE 11th Data Driven Control Learn. Syst. Conf. (DDCLS)*, Aug. 2022, pp. 391–395.

[21] J. Hu, P. Qiao, H. Lv, L. Yang, A. Ouyang, Y. He, and Y. Liu, "High speed railway fastener defect detection by using improved YoLoX-nano model," *Sensors*, vol. 22, no. 21, p. 8399, Nov. 2022.

[22] X. Wang, H. Gao, Z. Jia, and Z. Li, "BL-YOLOv8: An improved road defect detection model based on YOLOv8," *Sensors*, vol. 23, no. 20, p. 8361, Oct. 2023.

[23] X. Ni, H. Liu, Z. Ma, C. Wang, and J. Liu, "Detection for rail surface defects via partitioned edge feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5806–5822, Jun. 2022.

[24] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, Sep. 2022, pp. 443–459.

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

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[27] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.

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

[29] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[30] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9756–9765.

[31] Y. Zhou, W. Zhu, Y. He, and Y. Li, "YOLOv8-based spatial target part recognition," in *Proc. IEEE 3rd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, vol. 3, May 2023, pp. 1684–1687.

[32] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21002–21012.

[33] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000.

[34] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[35] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, "IoU loss for 2D/3D object detection," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 85–94.

[36] *Track Defect Detection Dataset*, SUSTECH, Roboflow, 2023. [Online]. Available: https://roboflow.com/

[37] *Clipping Dataset*, FENG, Roboflow, 2023. [Online]. Available: https://roboflow.com/

YAN WANG received the bachelor's degree in engineering from Zhejiang Normal University, Jinhua, China, in 2020, where she is currently pursuing the master's degree in engineering. Her current research interest includes machine vision.

KEHUA ZHANG received the B.E. degree from Dalian Jiaotong University, in 2000, the M.S. degree in engineering from Guangxi University, in 2004, and the Ph.D. degree in mechanical and electronic engineering from Zhejiang University of Technology, Hangzhou, China, in 2009. He is currently a Professor with Zhejiang Normal University. His research interests include slam, industrial robotics, machine learning, and computer vision.

LING WANG received the bachelor's degree in engineering from Zhejiang Normal University, Jinhua, China, in 2021, where he is currently pursuing the master's degree in engineering. His current research interest includes machine vision.

LINTONG WU received the B.E. degree from Zhejiang University of Water Resources and Electric Power, in 2021. He is currently pursuing the master's degree in electronic information with Zhejiang Normal University, Jinhua, China. His main research interest includes computer vision.