applied sciences

MDPI

Article

Lightweight Detection Method of Wheelset Tread Defects Based on Improved YOLOv7

Peng Yang [1,*], Fan Gao [1], Xinwen Yang [1], Caidong Wang [2], Hongjun Yang [1] and Zhifeng Zhang [1,*]

[1] School of Electronics and Information, Zhengzhou University of Light Industry, Zhengzhou 450002, China; yuegao17@outlook.com (F.G.); 060715@163.com (X.Y.); 2006014@zzuli.edu.cn (H.Y.)
[2] College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2011009@zzuli.edu.cn
* Correspondence: 2013005@zzuli.edu.cn (P.Y.); 2009041@zzuli.edu.cn (Z.Z.)

Abstract

Accurate online inspection of train wheelset tread defects is challenging owing to the variety and position uncertainty of defects. This study develops an improved YOLOv7 model capable of inspecting various wheelset tread defects with high accuracy and low computation complexity. This model comprises GSConv, a small target enhancement (STE) module, and StyleGAN3. GSConv significantly reduces the model volume while maintaining the feature expression ability, achieving a lightweight structure. The STE module enhances the fusion of shallow features and distribution of attention weights, significantly improving the sensitivity to small-sized defects and positioning robustness. StyleGAN3 enhances small samples by addressing inhomogeneity, thereby generating high-quality defect samples; it overcomes the limitations of traditional amplification methods regarding texture authenticity and morphological diversity, systematically improving the model's generalization ability under sample scarcity conditions. The model achieves 1.6%, 10.7%, 48.63% and 37.97% higher mean average precision values than YOLOv7, YOLOv5, SSD, and Faster R-CNN, respectively, and the model parameter size is reduced by 73.91, 94.69, 122.11, and 154.91 MB, respectively. Hence, the proposed YOLOv7-STE model outperforms traditional models. Moreover, it demonstrates satisfactory performance in detecting small target defects in different samples, highlighting its potential applicability in online wheel tread defect inspection.

Keywords: deep learning; small target detection; small target enhancement; wheelset tread defect; defect detection

Academic Editor: Douglas O'Shaughnessy

Received: 1 September 2025
Revised: 1 October 2025
Accepted: 6 October 2025
Published: 10 October 2025

Citation: Yang, P.; Gao, F.; Yang, X.; Wang, C.; Yang, H.; Zhang, Z. Lightweight Detection Method of Wheelset Tread Defects Based on Improved YOLOv7. *Appl. Sci.* **2025**, *15*, 10903. https://doi.org/10.3390/app152010903

1. Introduction

Train wheelsets, which are critical components of high-speed trains, are susceptible to failure under long-term operation in harsh environments [1,2]. The defects in wheelset treads typically include peels, wear, bruises, and pits. The efficient detection of wheelset-tread defects is necessary for the safe operation of trains [3,4]. Traditional methods primarily adopt manual measurements; however, most inspection information is not used appropriately for efficient detection [5]. Machine vision technology is frequently employed in tedious inspection tasks to reduce labor waste [6–8]. In complex environments, traditional machine vision inspection faces challenges such as vibration, noise, and light variation, which seriously influence measurement accuracy and efficiency. Numerous defect detection studies have focused on convolutional neural networks (CNNs) because of their superior

ability to overcome various disturbances [9]. The most common models mainly include two-stage models, such as faster region-based CNN (Faster R-CNN) models, for their accuracy [10,11], and two-stage models, such as You Only Look Once (YOLO), for their efficiency [12,13]. YOLO-based algorithms have been widely applied in numerous online and complex detection areas owing to their high efficiency [13–16].

An accurate and comprehensive online inspection of defects in wheelset treads is difficult because of their variety and position uncertainty; in particular, the accurate detection and location of small targets poses greater challenges. Compared to common targets, small targets usually have a low pixel count with a smaller portion and are easily overlapped; they are therefore difficult to recognize. In 2014 [17], Lin et al. first clearly defined the scales of small, medium and large targets based on pixel area and revealed that small targets are the main problem restricting detection accuracy, thus laying the foundation for subsequent research on small target detection. Small object detection, which aims to identify indistinguishable tiny objects in images, has been greatly advanced by Cheng et al. [18], who first provided a systematic survey of this field and constructed the large-scale dedicated benchmark SODA. Ren et al. [19] improved YOLOv4 to solve the problem of small target recognition by using a lightweight backbone and a shallow feature enhancement network. Tian et al. [20] proposed multiscale dense YOLO (MD-YOLO), which incorporates a feature extraction path and a feature aggregation path. A deep network, which leverages spatial location information from a shallower network, results in enhanced accuracy of small target detection. Ju, Luo and others used an improved deep learning network for feature extraction and small object detection. However, the detection speed and accuracy are still limited and cannot meet the actual detection accuracy requirements [21]. Zhang and Sun [22] embedded an attention module in the traditional network structure, which enhanced the feature extraction ability of the model. However, the method lacks universality and cannot meet the lightweight requirements. Cai et al. [23] achieved a more precise and efficient detection method that utilizes channel attention and spatial attention modules to suppress insignificant features in datasets. Kang et al. [24] proposed ASF-YOLO, which combines spatial and scale features for accurate and fast cell-instance segmentation. Zhou and Yu [25] significantly reduced the model size and ensured the detection accuracy, but the yolov5 model they used was too outdated and was prone to limitations in practical operation. Zhang Chen et al. [26]. innovatively proposed an early warning method for railway safety, which detects and extracts the track area based on the YOLOv5 model. However, the experiment lacks real data support and a practical verification effect.

This study proposes an improved YOLOv7 model that realizes the efficient detection of wheelset tread defects in high-speed trains. The captured images of the wheelset tread are preprocessed for small and nonuniform samples. First, the wheelset tread image dataset is expanded using traditional data augmentation and Copy–Paste [27] to address data scarcity. Second, the obtained dataset is trained using the StyleGAN3 network [28]. Third, the defects are labeled using the LabelImg software. The dataset is then divided into training, validation, and test sets. Finally, lightweight convolution in the Neck module is employed to reduce the computational parameters of the model. In addition, a small object enhancement (STE) module is introduced to enhance the multiscale feature expression ability by fusing feature maps of different scales using convolution and pooling operations. An attention mechanism is embedded in this module to suppress noise while highlighting the key features of minor defects, thereby significantly improving the model's sensitivity and discrimination ability for low-pixel, small objects. A comparison of the experimental results of YOLOv7-STE with YOLOv7, YOLOv5, SSD, and Faster R-CNN validated its superior detection accuracy. A publicly available rail surface defect dataset (RSDD) with similar small target defects was used to verify the robustness of the proposed model.

Compared to the traditional network models, the proposed model demonstrated good performance in detecting small target defects in different samples. The study findings confirm the potential of YOLOv7-STE for use in the online inspection of wheel tread defects.

2. Materials and Methods

2.1. Dataset, Environment, and Parameters

The representational ability of deep models is highly dependent on the diversity of training data. However, a fully publicly available dataset for wheelset damage detection does not currently exist. Therefore, the creation and expansion of such datasets is necessary for effective model training in deep learning-based wheelset damage detection.

To create the required dataset, images of various types of wheelset tread defects, as shown in Figure 1, were first captured. However, the number of useful defect images was limited and typically unbalanced. Therefore, preprocessing was performed to expand the sample size. The dataset was first expanded using traditional data augmentation and Copy–Paste; the results are shown in Figure 2. A dataset comprising 654 images was thus obtained; however, as it was insufficient, the StyleGAN3 network was introduced and trained. Figure 3 shows the wheelset tread defect images generated using StyleGAN3. To address imbalances in the dataset, StyleGAN3 was used to selectively generate the corresponding defective images for dataset expansion, and a new dataset comprising 1200 images was obtained. The defects were labelled using LabelImg software, and the dataset was divided into training, validation, and test sets in an 8:1:1 ratio. The training set was further expanded to 4800 images by splicing and blending the augmented data with Gaussian blur, affine transform, luminance transform, descending pixel transform, and flip transform [29,30]. The statistical information for the dataset is presented in Table 1. There were significant dimensional differences among the different defects, as shown in Figure 4.

Figure 1. Image of wheelset tread defects. There are three main types of defects—pit, bruise, and peel.

Figure 2. Extending dataset using traditional methods.

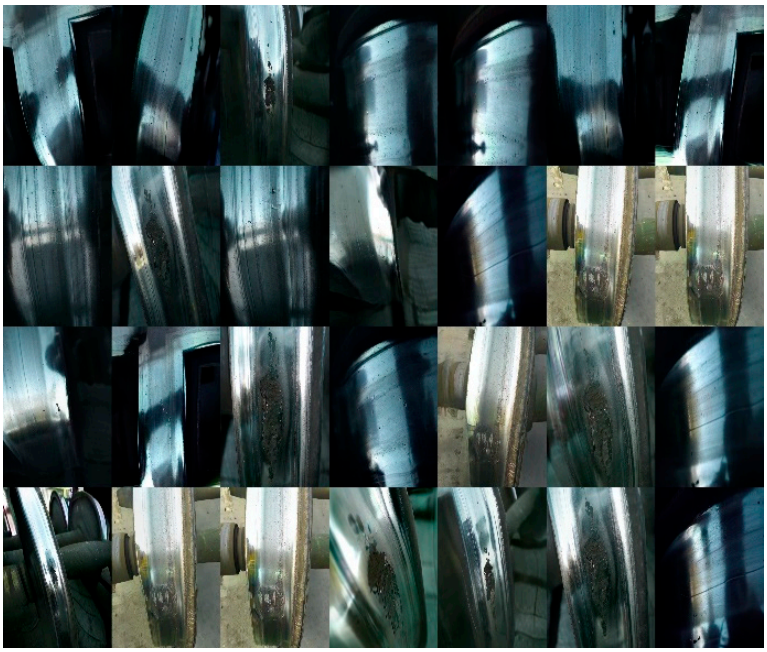

Figure 3. Wheel tread defect images generated by StyleGAN3.

Table 1. Target statistics of wheelset defects.

Categories	Number of Targets		
	Training Set	Validation Set	Test Set
Pit	1280	160	160
Bruise	1280	160	160
Peel	1280	160	160
Total	3840	480	480

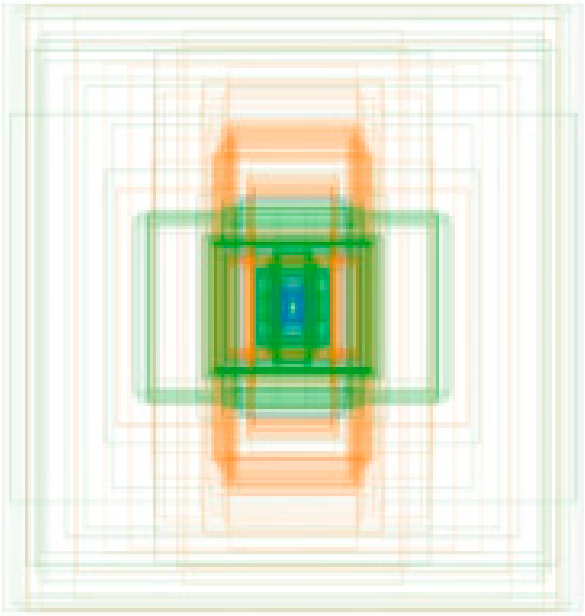

Figure 4. Size of each defect type in the annotation box (blue: pit; green: bruise; yellow: peel).

To quantitatively evaluate the quality of the images generated by StyleGAN3 and their consistency with the feature distribution in real defect images, the Frechet Inception Distance (FID) was calculated [31]. FID evaluates the generation quality by comparing the feature distribution distances of two types of images in the Inception-v3 feature space; the lower the value, the higher the visual fidelity and diversity of the generated image. The calculated FID value between the generated image and the real image was 15.3. According to relevant research, this FID value is relatively low, indicating that the high-quality samples generated by StyleGAN3 are significantly closer in terms of statistical features to the real defect images. This effectively demonstrates the reliability and effectiveness of the data augmentation method adopted in this study.

As shown in Figure 1, the acquired images mainly included three types of defects—peels, pits, and bruises. Pits are the most typical early initial defects. Although the features are minor, they can lead to the occurrence and deterioration of cracks. Pits occupy only a small number of pixels in the image and are typically encountered in small object detection. Peels generally result from corrosion development in pits and form in the intermediate stage of the transition from early defects to severe defects. Bruises are sudden, acute injuries that can damage the overall integrity of the wheel structure and lead to more serious derivative defects such as peeling.

These three are the most common types of defects that frequently occur and damage the surface of railway wheels during daily operations, resulting in material fatigue and mechanical wear. Thus, detection of such early defects is crucial in practical applications and holds extreme safety significance. Furthermore, the detection of such defects validates the innovative detection methods proposed in this study, such as small target enhancement and lightweighting. It provides good representativeness.

The hardware environment and software versions used in the experiments are listed in Table 2. In this study, the stochastic gradient descent (SGD) method was used to optimize the learning rate, and the epochs were determined by comparing the loss functions of the training and validation sets. The parameters of the training network are listed in Table 3. As shown in Table 3, the batch size is set to 64 to balance GPU memory utilization and gradient stability. The initial learning rate is 0.01, combined with a warm-up mechanism of 3 epochs to quickly escape the local optimal solution. Based on the 2160 target of the training set sample size, the number of iterations is set to 120, and the momentum coefficient is 0.937 for

accelerated convergence. The image size is set to 640 × 640 to ensure defect detection accuracy while optimizing computational efficiency. The optimizer adopts SGD to enhance model generalization and adapt to industrial deployment requirements.

Table 2. Experimental environment configuration.

Hardware and Software	Configuration Parameter
Computer	Operating System: Windows 10
	CPU: Intel(R) Core (TM) i9-9900K CPU@3.60GHz
	GPU: NVIDIA GeForce RTX 3090
	RAM: 16 GB
	Video memory: 24 GB
Software version	Python3.9.12 + PyTorch1.9.1 + CUDA11.7 + cuDNN8.2.1 + Opencv4.5.5 + Visual Studio Code2022 (1.69.1)

Table 3. Training network parameters.

Parameter	Value
Batch size	64
Learning rate	0.01
Warm-up epochs	3
Number of iterations	120
Momentum parameter	0.937
Image size	640 × 640
Optimizer	SGD

2.2. Loss Function and Model Evaluation Metrics

To verify the superior performance of the improved YOLOv7 model, we determined the mean average precision (mAP), frames per second (*FPS*), and model volume. The commonly used metrics of precision (*P*), recall (*R*), and mAP were selected to evaluate model performance [32]; these metrics are defined as follows:

$$P_{mAP} = \frac{1}{C_{class}} \sum_{i=1}^{C_{class}} \int_0^1 P(R) dR \tag{1}$$

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{2}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{3}$$

Here, *TP* denotes positive samples predicted to be correct (i.e., true positives), *FP* denotes negative samples predicted to be incorrect (i.e., false positives), *FN* denotes positive samples predicted to be incorrect (i.e., false negatives), and *N* denotes the number of sample categories.

The loss function of YOLOv7 comprises the following three components: confidence, bounding-box regression, and classification losses. The YOLOv7 loss function can be expressed as follows:

$$Loss_{total} = \lambda_1 L_{obj} + \lambda_2 L_{box} + \lambda_3 L_{cls}, \tag{4}$$

where L_{obj}, L_{box}, and L_{cls} denote confidence, bounding-box regression, and classification losses, respectively. Furthermore, λ_1, λ_2, and λ_3 are weight coefficients for the three losses;

changing these coefficients can adjust the emphasis on the three losses. In YOLOv7, L_{box} is calculated using the complete intersection over union (CIoU) loss, L_{CIoU} [33], which improves the speed and accuracy of bounding box regression. The expression for L_{CIoU} is as follows:

$$L_{CIOU} = 1 - IOU + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha v, \tag{5}$$

$$IOU = \frac{|b \cap b_{gt}|}{|b \cup b_{gt}|}, \tag{6}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w}{h}\right)^2. \tag{7}$$

Here, b and b_{gt} denote the predicted box and ground truth box, respectively; w_{gt}, h_{gt}, w, and h denote the width and height of the ground truth box and predicted box, respectively; ρ represents the distance between the centers of the two boxes; c represents the maximum distance between the boundaries of the two boxes; and α denotes a weight coefficient.

2.3. Improved YOLOv7 Network Architecture

2.3.1. YOLOv7 Network Architecture

YOLOv7 is a widely used detection network and is suitable for scenes requiring high detection accuracy. The network structure comprises Head, Neck, and Backbone modules as shown in Figure 5. There are three common variants of the YOLOv7 model—YOLOv7, YOLOv7x, and YOLOv7tiny. YOLOv7tiny is the smallest model with a fast detection speed and relatively low accuracy, and it is mainly applied to scenes with high speed requirements or limited computational resources. In this study, YOLOv7 was used as the base model, and a few parameters were adjusted to improve the model's performance.

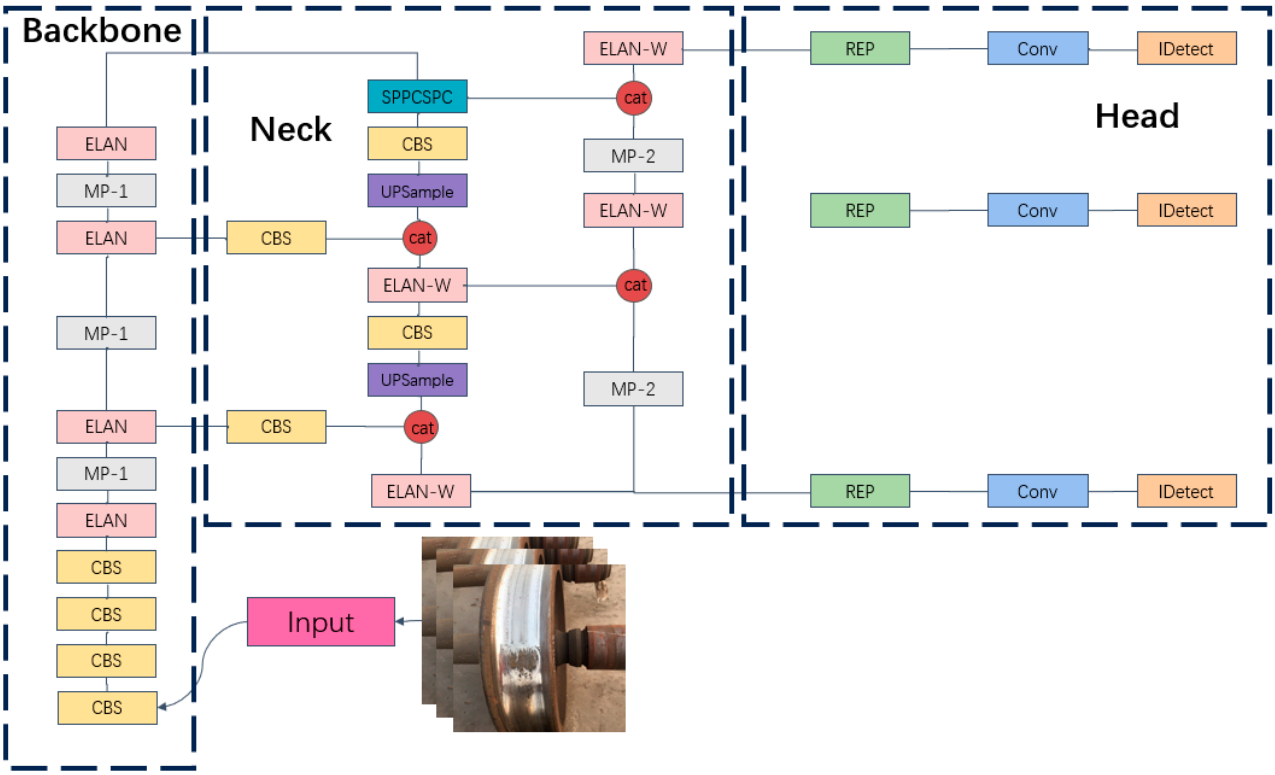

Figure 5. Structure of YOLOv7 model.

2.3.2. YOLOv7-STE

The Neck module of YOLOv7 was improved to enhance the detection and classification of different types of wheelset tread defects. GSConv was introduced instead of Conv to guarantee detection accuracy, reduce the network volume and computation amount, and obtain a lightweight design of the model. To detect and recognize small pits in the defect images, an STE module was added, which performed feature fusion of three different scales of feature maps (p3, p4, and p5) in the Backbone module to capture multiscale features. The structure of the YOLOv7-STE network is illustrated in Figure 6.

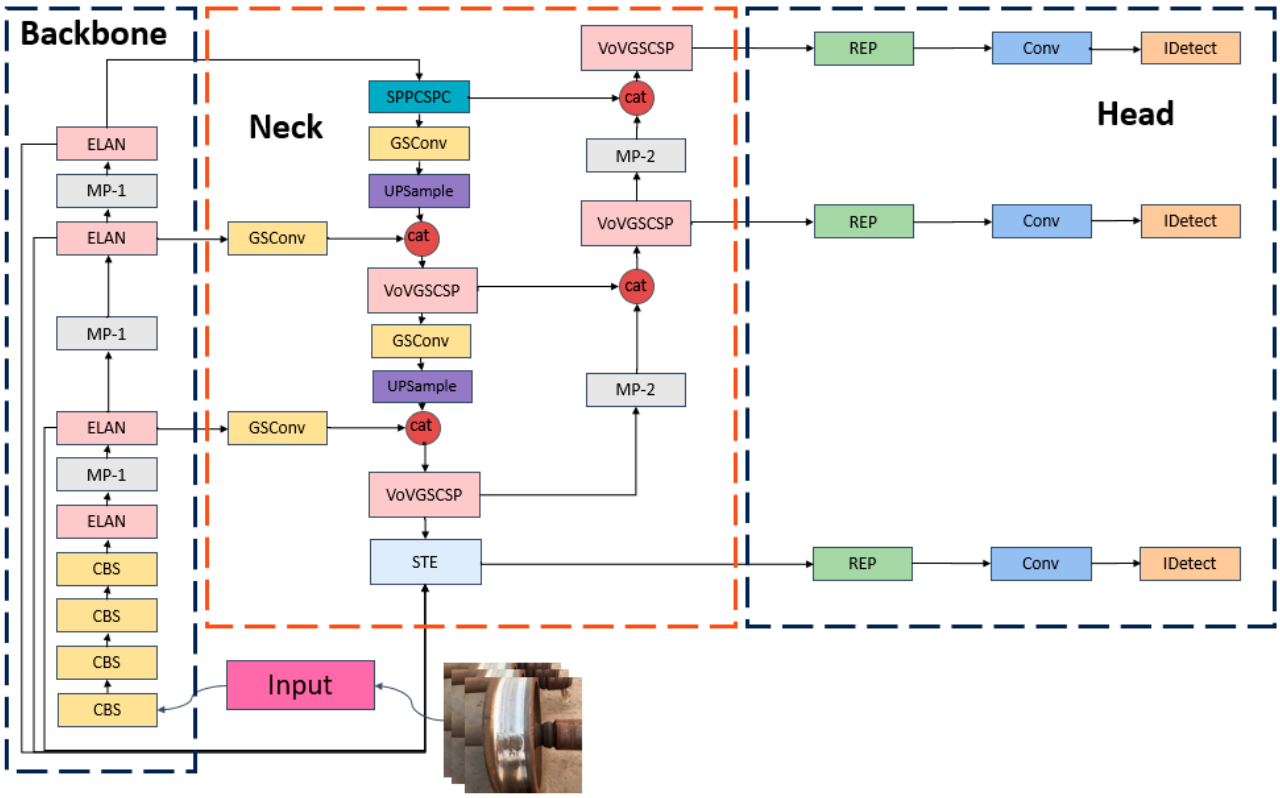

Figure 6. Improved YOLOv7 model.

The multiscale module primarily performs the fusion of features from different scales in the Backbone module via 3D convolution and pooling; in particular, it provides better feature representation for the detection of small targets. Furthermore, channel-attention and spatial-attention modules are introduced for multiscale features to fully use cross-channel information and achieve improved performance without increasing network complexity, as shown in Figure 7. By introducing channel attention and spatial attention mechanisms, the multiscale features extracted by the convolutional neural network are weighted; this weakens the background noise and non-significant regions or feature responses unrelated to defects, while enhancing the key features related to wheel tread defects. This processing helps to enhance the model's ability to detect small target defects and improve the model's generalization performance and robustness.

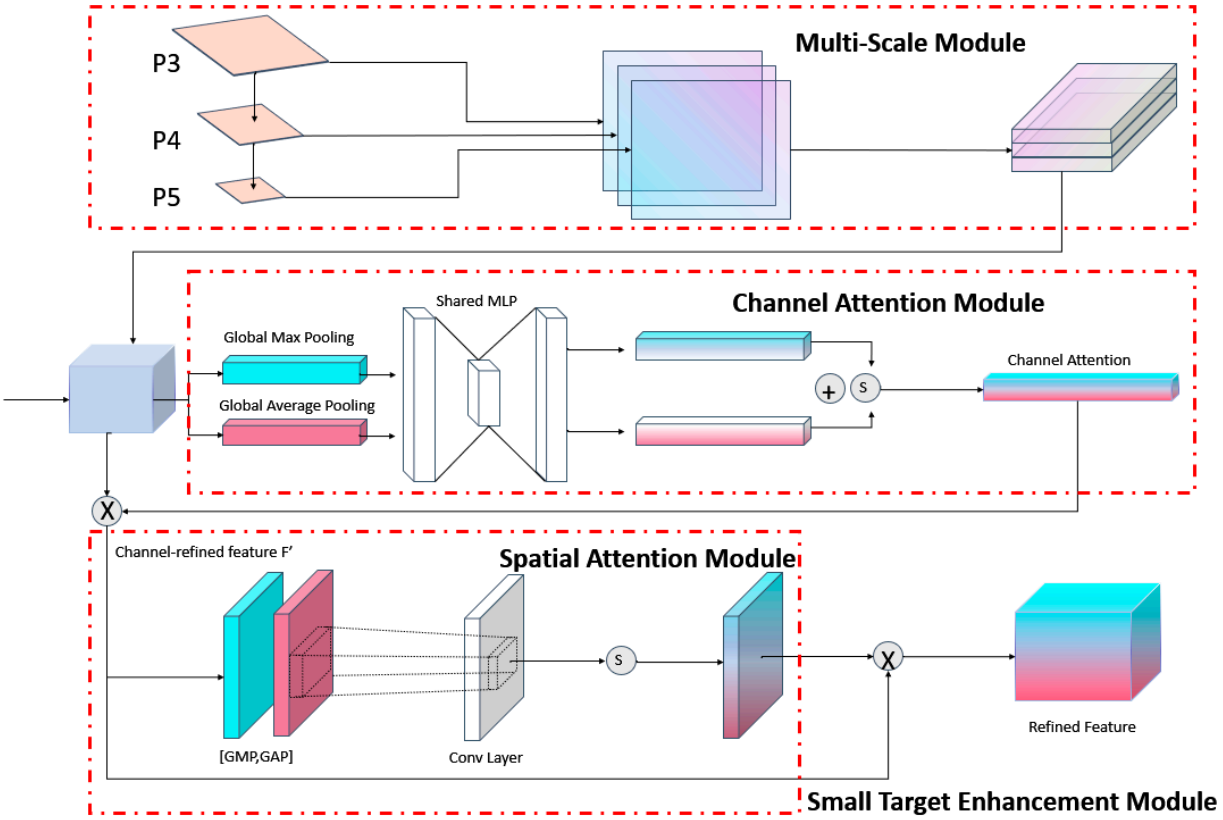

Figure 7. Small target enhancement module.

2.3.3. Loss Function

CIoU is sensitive to aspect ratio variations; thus, aspect ratio matching is generally overemphasized at the expense of other factors such as the overall positional relationship of the target. Therefore, enhanced intersection over union (EIoU) [34] is mainly used in target detection to evaluate the difference between predicted and real bounding boxes. The expression for EIoU is presented in Equation (11).

Given that the distance from the center is D_c, the center of the prediction box is (x_p, y_p), the true box center is (x_g, y_g), the width difference is w_{diff}, and the height difference is h_{diff}. The following relationships are obtained:

$$D_c = \sqrt{(x_p - x_g)^2 + (y_p - y_g)^2}, \tag{8}$$

$$w_{diff} = |w_p - w_g|, \tag{9}$$

$$h_{diff} = |h_p - h_g|, \tag{10}$$

$$EIoU = IoU - \lambda_1 \frac{D_C^2}{C_d^2} - \lambda_2 \frac{w_{diff}^2}{C_w^2} - \lambda_3 \frac{h_{diff}^2}{C_h^2} \tag{11}$$

where λ_1, λ_2, and λ_3 are hyperparameters that regulate the effects of center distance and width–height differences; and C_d, C_w, and C_h are constants used for normalization, and they are related to the size of the target in the dataset.

The utilization of EIoU provides several advantages. First, the comprehensive consideration of multiple geometric information of the bounding box, including overlapping area, center distance, and width–height difference, enables the model to more accurately learn the position and shape information of the target. This improves the target detection accuracy, particularly in complex situations such as deformation and occlusion of the target.

Second, compared to some traditional loss functions, the width and height loss of EIoU directly minimizes the difference between the width and height of the target and prediction frames; this can guide the model to quickly learn the appropriate bounding box parameters, accelerating the convergence speed of the model. Third, the EIoU loss function enhances the small target bounding box.

2.4. Discussion on the Innovations in the Improved YOLOv7 Model

To address the challenges in wheelset defect detection—small targets and low pixels—as well as the model lightweighting and real-time performance requirements in industrial scenarios, this study systematically improved the YOLOv7 network. First, to optimize the model efficiency and reduce the number of parameters, the GSConv module was introduced in the Neck part. This structure was initially introduced by Chen et al. [35] to significantly reduce model computational complexity while maintaining accuracy; it is particularly suitable for resource-constrained embedded deployment environments. Second, to enhance the recognition ability of minor defects, the STE module was introduced by integrating multiscale features and attention mechanisms; it is inspired by the idea introduced by Huillcen et al. [36] in agricultural droplet detection to improve the sensitivity of small targets through feature enhancement. Finally, in response to the accuracy requirements of the regression of the defect bounding box of the wheel pair, the EIoU loss function was adopted to replace the standard CIoU. Liu et al. [37] pointed out in pavement crack detection that EIoU can more directly optimize bounding box regression by decoupling the width–height loss term, which is particularly beneficial for small target positioning. The innovation of this work lies in systematically integrating and adapting these strategies that have been proven effective in different visual tasks to the specific scenario of wheelset defect detection, ultimately achieving a good balance between light weight and detection accuracy.

3. Experimental Results and Discussion

3.1. Comparative Experimental Results Analysis

The experimental results of the proposed model and the mainstream detection algorithms are compared in Table 4. Compared with the YOLOv7, YOLOv5, SSD, and Faster R-CNN models, the proposed model achieved 1.2%, 9.8%, 47.1%, and 38.7% higher mAP values, respectively, and the model parameter size was lower by 73.91, 94.69, 122.11, and 154.91, respectively. Moreover, compared with YOLOv8, the proposed model had a 1.6% higher mAP, and its parameter size was reduced by 2.01 MB. The proposed model exhibited better detection accuracy than the other models. Table 4 highlights the significantly better overall performance of the proposed model. For a more comprehensive performance evaluation, the F1 scores of each comparison model in different defect categories were obtained (Table 5). This indicator integrates accuracy and recall rate, and it evenly reflects the model's comprehensive recognition ability for various categories.

Table 4. Comparison of various detection models.

Model	Parameter size (MB)	mAP@0.5 (%)	mAP@0.5:0.95(%)	FPS
YOLOv7-STE	61.09	98.1	65.3	75.9
YOLOv8	63.1	96.5	56.8	81.1
YOLOv7	135	96.9	55.2	75.4
YOLOv5	155.78	88.5	51.2	73.6
SSD	183.2	51.9	40.7	33.1
Faster R-CNN	216	60.5	35.9	8.3

Table 5. Comparison of F1 scores (%) of different detection models for various defects.

Model	Pit	Bruise	Peel	Macro-Avg
YOLOv7-STE	97.3	97.5	99.6	98.1
YOLOv7	86.5	94.1	98.8	93.1
YOLOv8	87.8	94.5	98.9	93.7
YOLOv5	80.1	89.3	95.2	88.2

Based on the experimental results of this study, a paired *t*-test was conducted to evaluate the statistical significance of the performance improvement of YOLOv7-STE compared to that of its baseline, YOLOv7. The evaluation used mAP@0.5 values under a 50% verification discount, with the following paired results: YOLOv7 = [94.8, 95.2, 95.5, 94.6, 95.1], yolov7-ste = [95.9, 96.3, 96.5, 95.8, 96.2]. The t-statistic value was calculated to be 3.12, and the *p*-value was 0.035 ($p < 0.05$). The results show that although the performance improvement is limited, it is statistically significant, indicating that the improved strategy presented a stable and beyond random fluctuations impact. This finding provides preliminary statistical evidence for the validity of the model and suggests that future research needs to further optimize the model structure or training strategy to achieve more significant performance improvements.

3.2. Detailed Error Analysis and Security Discussion

Although the aforementioned results indicate that the YOLOv7-STE model exhibits excellent overall performance for all indicators, its ability to detect different types of defects may vary. For a comprehensive assessment of the practical application potential of the model and to identify its potential weak links, a fine-grained performance decomposition of the detection results of various defects is presented in Table 6.

Table 6. Detailed performance evaluation of the YOLOv7-STE model by defect type.

Defect Category	Precision	Recall	AP@0.5	False Negative Rate
Pit	92.1	92.5	90.2	7.5
Bruise	96.8	95.9	95.3	4.1
Peel	99.2	98.8	99.0	1.2
All categories	96.0	95.7	94.8	4.3

The analysis shows that although the overall performance of the model is excellent, its detection ability varies for different types of defects. For large-sized defects such as "Peel" and "Bruise", the model performs exceptionally well with an extremely low false negative rate. However, for the "Pit" defect with a small pixel ratio and indistinct features, its false negative rate is significantly higher than that of other categories. Thus, the model is currently limited in terms of its reliability in detecting extremely small target defects.

3.3. Ablation Experiments

Ablation experiments were conducted to explore the impact of the proposed improvements on the network model. The results are presented in Tables 7 and 8. We conducted eight groups of experiments, each with different modules, and mAP@0.5, model volume, mAP@0.5:0.95, and FPS as indicators. The results were compared with those of the original YOLOv7 model. Of the five different loss functions used on the original YOLOv7 model, the EIoU loss yielded the best performance (Table 7); mAP@0.5 increased by 1.8% and mAP@0.5:0.95 increased by 1.44% as compared to the original YOLOv7 model (with CIoU loss). The EIoU loss function therefore improves the effect of small-object bounding boxes.

Table 7. Comparison of different IoUs.

Loss Function	Model Volume (MB)	mAP@0.5:0.95 (%)	mAP@0.5(%)			
			All Classes	Pit	Peel	Bruise
CIoU	135	51.33	93.9	91.6	97.3	92.9
WIoU	135	51.44	94.1	90	99.5	93
SIoU	135	51.34	93.9	91.5	97.3	93
DIoU	135	51.69	94.6	91.1	99.2	93.6
EIoU	135	52.77	95.7	92.6	99.5	95.1

Table 8. Results of ablation experiments.

GSConv	STE	EIoU	Model Volume (MB)	mAP@0.5:0.95 (%)	mAP@0.5(%)				FPS
					All Classes	Pit	Peel	Bruise	
		√	135	52.9	96.0	92.8	99.5	95.8	73.9
√		√	51	50.3	92.7	83.3	99.6	95.5	86.7
	√	√	149	53.1	96.1	97.1	97.1	94.1	61
√	√	√	61.09	65.3	98.1	97.3	99.6	97.5	75.9

Table 8 shows that the model volume of the GSConv layer in the Neck module reduced by 84 MB, whereas the FPS increased to 87.1. Regarding small target (Pit) detection, mAP@0.5 reduced by 10.2%. The introduction of the STE module improved the small-target detection mAP by 4.4%, whereas the model volume increased by 14 MB. The detection speed decreased, and the mAP was lower for the rest of the target sizes, with a 2.6% reduction in mAP@0.5 for Peel and a 1.3% reduction for Bruise. When both the STE module and GSConv were used, compared to YOLOv7 (EIoU), the model volume was reduced by 73.91 MB. The mAP@0.5 (all classes) improved by 1.6%, with a 4.2% improvement in the mAP of Pit, 0.1% in the mAP of Peel, 0.3% in the mAP of Bruise, and 1.9% in FPS. In summary, the improved strategy based on YOLOv7 proposed in this study significantly improves wheelset tread damage detection.

3.4. Comparison of Different Models by Defect Type

To verify the superiority of the proposed YOLOv7-STE model for wheelset tread damage detection, the results were compared with those of YOLOv7, YOLOv5, SSD, and Faster R-CNN models. The relevant parameters of the experiments were strictly controlled using a uniform image size as the input and a uniform training and test set for experimental testing. The experimental dataset used is detailed in Table 1; the control environment and parameter settings for the experiments were the same. A comparison of the recognition of pits using the different models is shown in Figure 8. The confidence levels of detection results of YOLOv7-STE, YOLOv7, YOLOv5, SSD, and Faster R-CNN were 0.96, 0.86, 0.77, 0.69, and 0.58, respectively. The recognition of bruises using the different models is shown in Figure 9. The confidence levels of YOLOv7-STE, YOLOv7, YOLOv5, SSD, and Faster R-CNN detection results were 0.95, 0.91, 0.92, 0.85, and 0.77, respectively. Figure 10 shows a comparison of the recognition results for Peel by different models. The confidence levels of YOLOv7-STE, YOLOv7, YOLOv5, SSD, and Faster R-CNN detection results were 0.97, 0.95, 0.95, 0.90, and 0.89, respectively. The experimental results indicate that the proposed method was more adaptable to complex backgrounds than the other models and achieved better inspection performance.

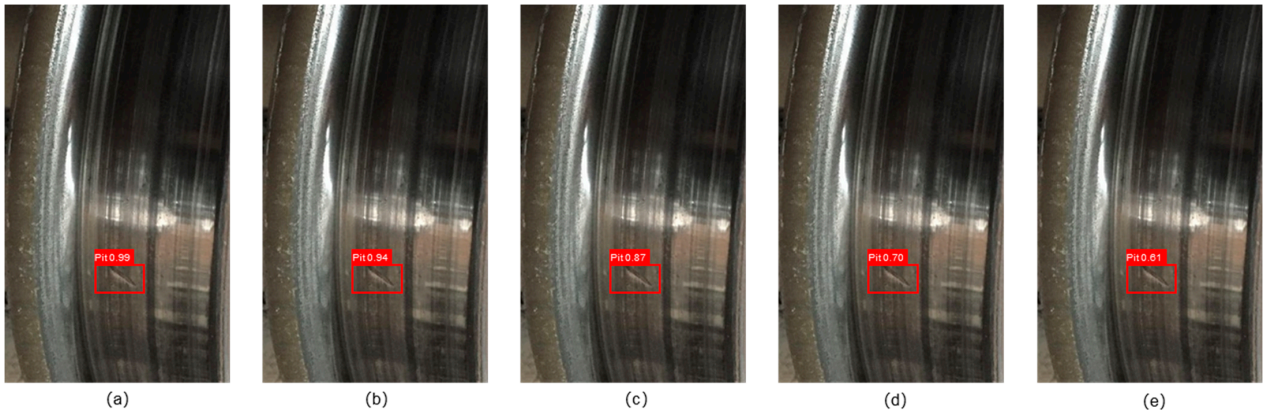

Figure 8. Comparison of the recognition of pits using different models. (**a**) Confidence level of YOLOv7-STE detection results. (**b**) YOLOv7 detection results. (**c**) YOLOv5 detection results. (**d**) SSD detection results. (**e**) Faster R-CNN detection results.

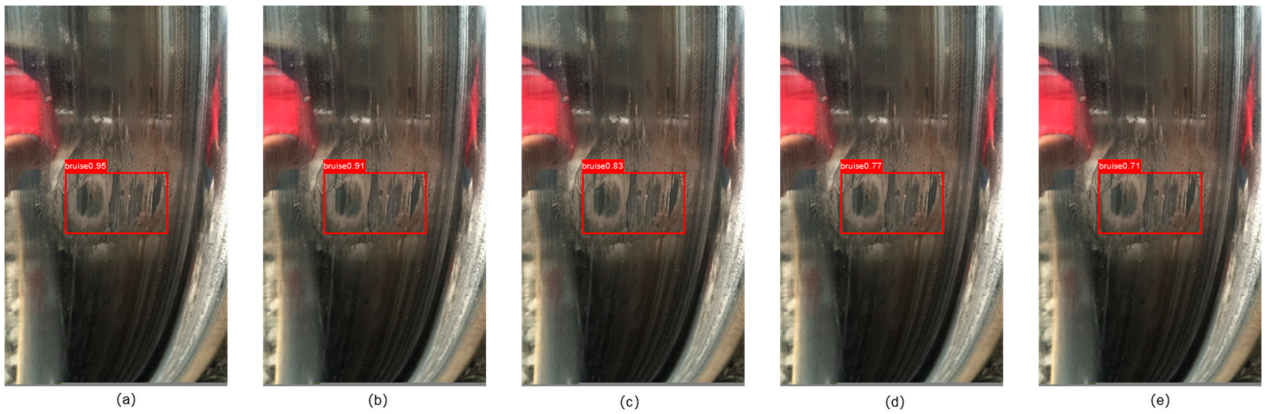

Figure 9. Comparison of the recognition of bruise using different models. (**a**) Confidence level of YOLOv7-STE detection results. (**b**) YOLOv7 detection results. (**c**) YOLOv5 detection results. (**d**) SSD detection results. (**e**) Faster R-CNN detection results.

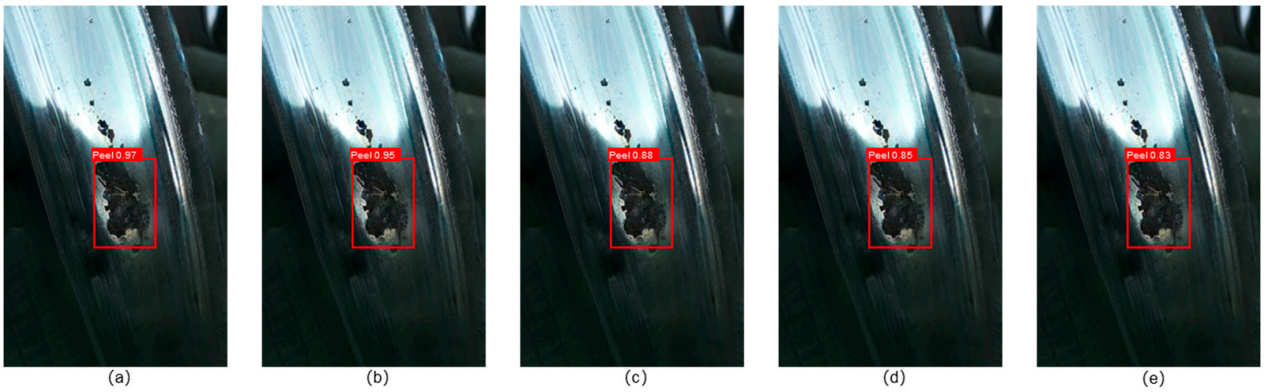

Figure 10. Comparison of the recognition of peel using different models. (**a**) Confidence level of YOLOv7-STE detection results. (**b**) YOLOv7 detection results. (**c**) YOLOv5 detection results. (**d**) SSD detection results. (**e**) Faster R-CNN detection results.

A publicly available RSDDs dataset with similarly small target defects was used for verifying the robustness of the model. The experimental results are presented in Figure 11. The confidence level of YOLOv7-STE is remarkably high, at 0.97. Furthermore, the confidence levels of YOLOv7, YOLOv5, SSD, and Faster-RCNN are 0.90, 0.88, 0.52, and 0.49, respectively.

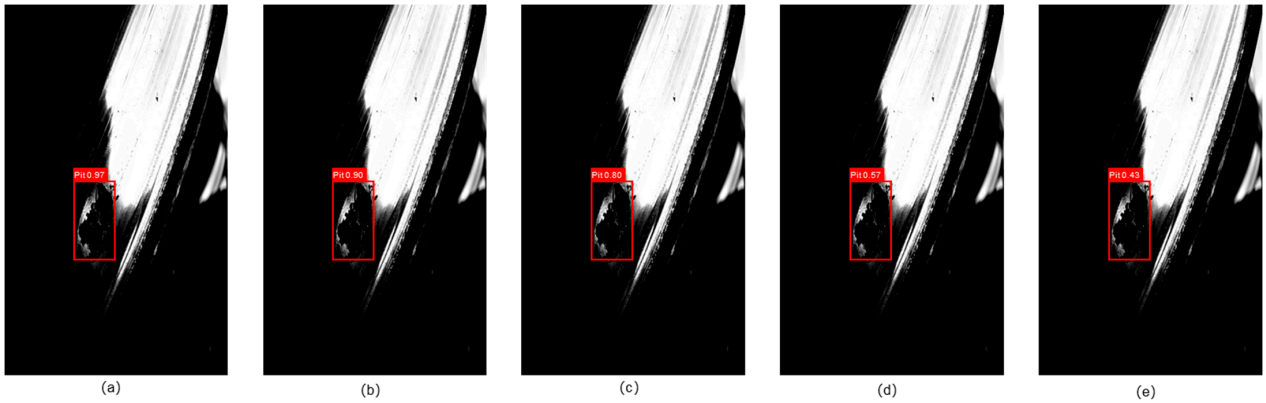

Figure 11. Different models are used to detect RSDDs in datasets. (**a**) Confidence level of YOLOv7-STE detection results. (**b**) YOLOv7 detection results. (**c**) YOLOv5 detection results. (**d**) SSD detection results. (**e**) Faster-RCNN detection results.

Defect detection results showed that, for large-sized defects (Peel and Bruise), the five models were effective, with YOLOv7-STE exhibiting the highest correct classification rate. For small defects (Pit), YOLOv7-STE exhibited the highest recognition rate and correct classification rate. In summary, the experimental results confirm that the YOLOv7-STE model outperformed the other four models. When tested on the RSDD, YOLOv7-STE demonstrated superior performance compared to traditional deep neural networks. The proposed model showed remarkable robustness to small target defects in different samples.

4. Conclusions

This study developed an improved YOLOv7 model for the inspection of wheelset tread defects, offering improved accuracy and reduced computation complexity. GSConv, a lightweight convolution layer, was integrated into the Neck module to reduce the model volume. The STE module addressed the challenge of low pixels and difficulty in distinguishing small targets, thereby improving the classification accuracy of damage feature recognition. The captured images of the wheelset tread were preprocessed to augment small and nonuniform samples through Copy–Paste, followed by training with the StyleGAN3 network. A comparison of the experimental results showed that the mAP value of the proposed model was higher by 1.2%, 9.8%, 47.1%, and 38.7% compared with those of YOLOv7, YOLOv5, SSD, and Faster R-CNN models, respectively, and it surpassed YOLOv8 by 1.6%. The model parameter size was lower by 73.91 MB, 94.69 MB, 122.11 MB, and 154.91 MB, respectively, when compared to those of the aforementioned detection models, and it was reduced by 2.01 MB compared to that of YOLOv8. The YOLOv7-STE model exhibited a significantly better overall performance than the traditional models. Its robustness was verified using the publicly available dataset, RSDD, which comprises similar small target defects. Compared to the traditional models, YOLOv7-STE demonstrated better performance for small target defects in different samples. The experimental results confirmed that the YOLOv7-STE model can satisfy the requirements of online inspection of wheelset tread defects.

However, in this study, owing to the limited collection of defects data, only three common types of damage (Pit, Bruise, and Peel) were identified and classified. In real scenarios, owing to the influence of a variety of complex factors (including differences in temperature and light intensity), similar defects (or even the same defects in different spatial and temporal environments) can lead to differences in manifestation. Hence, it is necessary to further expand the dataset and conduct different types of classification and identification tasks.

Although this study demonstrated high accuracy and FPS in the experimental environment, there is uncertainty about whether its computational efficiency and lightweight structure meet the requirements of embedded devices. Future research will focus on enhancing the applicability and practicality of the model in real and complex operating environments. This includes developing lightweight deployment solutions for real-time detection and optimizing the inference efficiency of the model on embedded devices to meet the requirements of online detection under high-speed operating conditions. For actual interference factors, such as drastic changes in light, rain and snow pollution, and motion blur, it is necessary to construct a more robust, adaptive detection model by introducing composite defect categories, conducting in-depth research on recognition methods based on few-shot learning, and enhancing the generalization ability of the model.

Author Contributions: Conceptualization: P.Y., Z.Z. and C.W.; Methodology: P.Y., X.Y. and H.Y.; Writing—original draft: F.G., X.Y. and H.Y.; writing—review and editing: F.G., C.W. and H.Y.; validation: P.Y. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Department of Science and Technology of Henan Province (242102211064), Key Research Projects of Higher Education Institutions in Henan Province (23B416001), Key Research and Development Special Projects in Henan Province (241111230300), Henan Major Science and Technology Project (221100230100), and the Henan Natural Science Foundation (242300420294).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, H.; He, J.; Liu, Z.; Zhang, C. LLD-MFCOS: A Multiscale anchor-free detector based on label localization distillation for wheelset tread defect detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 500385. [CrossRef]
2. Song, Y.; Ji, Z.; Guo, X.; Hsu, Y.; Feng, Q.; Yin, S. A comprehensive laser image dataset for real-time measurement of wheelset geometric parameters. *Sci. Data* **2024**, *11*, 462. [CrossRef]
3. Liu, J.; Jiang, S.; Wang, Z.; Liu, J. Detection of Train Wheelset Tread Defects with Small Samples Based on Local Inference Constraint Network. *Electronics* **2024**, *13*, 2201. [CrossRef]
4. Zhang, C.; Xu, Y.; Yin, H.L. Deformable residual attention network for defect detection of train wheelset tread. *Vis. Comput.* **2024**, *40*, 1775–1785. [CrossRef]
5. Zhang, X.; Lu, J. Integrated intelligent system for rail flaw detection vehicle. *Electr. Drive Locomot.* **2021**, *1*, 133–137. [CrossRef]
6. Luo, J.; Yu, X.; Cao, J.; Du, W. Intelligent rail flaw detection system based on deep learning and support vector machine. *Electr. Drive Locomot.* **2021**, *2*, 100–107. [CrossRef]
7. Wang, X.; Fu, Z. A new method of wavelet and support vector machine for detection of the train wheel bruise. *China Mech. Eng.* **2004**, *15*, 1641–1643.
8. Xu, Z.; Chen, J. Tread profile of wheel detection method based image processing and Hough transform. *Electron. Meas. Technol.* **2017**, *40*, 117–121.
9. Sheng, Z.; Wang, G. Fast method of detecting packaging bottle defects based on ECA-EfficientDet. *J. Sens.* **2022**, *2022*, 9518910. [CrossRef]
10. Palazzetti, L.; Rangarajan, A.K.; Dinca, A.; Boom, B.; Popescu, D.; Offermans, P.; Pinotti, C.M. The hawk eye scan: *Halyomorpha halys* detection relying on aerial tele photos and neural networks. *Comput. Electron. Agric.* **2024**, *226*, 109365. [CrossRef]
11. Shi, C.H.; Yang, H.F.; Cai, J.H.; Zhou, L.C.; He, Y.T.; Su, M.H.; Zhao, X.-J.; Xun, Y.-L. A Survey of Galaxy Pairs in the SDSS Photometric Images based on Faster-RCNN. *Astron. J.* **2024**, *168*, 90. [CrossRef]
12. Guo, H.; Wu, T.; Gao, G.; Qiu, Z.; Chen, H. Lightweight safflower cluster detection based on YOLOv5. *Sci. Rep.* **2024**, *14*, 18579. [CrossRef]
13. Li, Z.; Zhu, Y.; Sui, S.; Zhao, Y.; Liu, P.; Li, X. Real-time detection and counting of wheat ears based on improved YOLOv7. *Comput. Electron. Agric.* **2024**, *218*, 108670. [CrossRef]
14. Hsieh, C.C.; Hsu, C.H.; Huang, W.H. A two-stage road sign detection and text recognition system based on YOLOv7. *Internet Things* **2024**, *27*, 101330. [CrossRef]

15. Wang, D.; Qian, Y.; Lu, J.; Wang, P.; Hu, Z.; Chai, Y. Fs-yolo: Fire-smoke detection based on improved YOLOv7. *Multimed. Syst.* **2024**, *30*, 215. [CrossRef]
16. Liu, Y.; Jiang, B.; He, H.; Chen, Z.; Xu, Z. Helmet wearing detection algorithm based on improved YOLOv5. *Sci. Rep.* **2024**, *14*, 8768. [CrossRef]
17. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2014.
18. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards large-scale small object detection: Survey and benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef]
19. Ren, Y.; Zhang, H.; Sun, H.; Ma, G.; Ren, J.; Yang, J. LightRay: Lightweight network for prohibited items detection in X-ray images during security inspection. *Comput. Electr. Eng.* **2022**, *103*, 108283. [CrossRef]
20. Tian, Y.; Wang, S.; Li, E.; Yang, G.; Liang, Z.; Tan, M. MD-YOLO: Multi-scale Dense YOLO for small target pest detection. *Comput. Electron. Agric.* **2023**, *213*, 108233. [CrossRef]
21. Ju, M.; Luo, J.; Liu, G.; Luo, H. A real-time small target detection network. *SIViP* **2021**, *15*, 1265–1273. [CrossRef]
22. Zhang, M.; Su, H.; Wen, J. Classification of flower image based on attention mechanism and multi-loss attention network. *Comput. Commun.* **2021**, *179*, 307–317. [CrossRef]
23. Cai, Y.; Yao, Z.; Jiang, H.; Qin, W.; Xiao, J.; Huang, X.; Pan, J.; Feng, H. Rapid detection of fish with SVC symptoms based on machine vision combined with a NAM-YOLO v7 hybrid model. *Aquaculture* **2024**, *582*, 740558. [CrossRef]
24. Kang, M.; Ting, C.M.; Ting, F.F.; Phan, R.C.W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [CrossRef]
25. Yuan, Z.; Yue, X. Lightweight object detection algorithm for automotive fuse boxes based on deep learning. *J. Electron. Imaging* **2025**, *34*, 013031. [CrossRef]
26. Zhang, Z.; Chen, P.; Huang, Y.; Dai, L.; Xu, F.; Hu, H. Railway obstacle intrusion warning mechanism integrating YOLO-based detection and risk assessment. *J. Ind. Inf. Integr.* **2024**, *38*, 100571. [CrossRef]
27. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2918–2928.
28. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 852–863.
29. Abayomi-Alli, O.O.; Damaševičius, R.; Misra, S.; Maskeliūnas, R. Cassava disease recognition from low-quality images using enhanced data augmentation model and deep learning. *Expert Syst.* **2021**, *38*, e12746. [CrossRef]
30. Ye, Y.; Li, Y.; Ouyang, R.; Zhang, Z.; Tang, Y.; Bai, S. Improving machine learning based phase and hardness prediction of high-entropy alloys by using Gaussian noise augmented data. *Comput. Mater. Sci.* **2023**, *223*, 112140. [CrossRef]
31. Wang, W.; Zhang, M.; Wu, Z.; Zhu, P.; Li, Y. Scgan: Semi-centralized generative adversarial network for image generation in distributed scenes. *Inf. Fusion* **2024**, *112*, 102556. [CrossRef]
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
33. Yang, J.; Zhang, X.; Song, C. Research on a small target object detection method for aerial photography based on improved YOLOv7. *Vis. Comput.* **2024**, *41*, 3487–3501. [CrossRef]
34. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
35. Shengde, C.; Junyu, L.; Xiaojie, X.; Jianzhou, G.; Shiyun, H.; Zhiyan, Z.; Yubin, L. Detection and tracking of agricultural spray droplets using GSConv-enhanced YOLOv5s and DeepSORT. *Comput. Electron. Agric.* **2025**, *235*, 110353. [CrossRef]
36. Huillcen Baca, H.A.; Palomino Valdivia, F.L.; Gutierrez Caceres, J.C. Efficient human violence recognition for surveillance in real time. *Sensors* **2024**, *24*, 668. [CrossRef]
37. Liu, Z.; Gu, X.; Chen, J.; Wang, D.; Chen, Y.; Wang, L. Automatic recognition of pavement cracks from combined GPR B-scan and C-scan images using multiscale feature fusion deep neural networks. *Autom. Constr.* **2023**, *146*, 104698. [CrossRef]