applied sciences

MDPI

Article

DBCW-YOLO: A Modified YOLOv5 for the Detection of Steel Surface Defects

Jianfeng Han *, Guoqing Cui, Zhiwei Li and Jingxuan Zhao

School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China; cuiguoqing@stumail.tjcu.edu.cn (G.C.); 17320096827@163.com (Z.L.); 18500895063@163.com (J.Z.)
* Correspondence: hanjianfeng@toec-gdgs.com

Abstract: In steel production, defect detection is crucial for preventing safety risks, and improving the accuracy of steel defect detection in industrial environments remains challenging due to the variable types of defects, cluttered backgrounds, low contrast, and noise interference. Therefore, this paper introduces a steel surface defect detection model, DBCW-YOLO, based on YOLOv5. Firstly, a new feature fusion strategy is proposed to optimize the feature map fusion pair model using the BiFPN method to fuse information at multiple scales, and CARAFE up-sampling is introduced to expand the sensory field of the network and make more effective use of the surrounding information. Secondly, the WIoU uses a dynamic non-monotonic focusing mechanism introduced in the loss function part to optimize the loss function and solve the problem of accuracy degradation due to sample inhomogeneity. This approach improves the learning ability of small target steel defects and accelerates network convergence. Finally, we use the dynamic heads in the network prediction phase. This improves the scale-aware, spatial-aware, and task-aware performance of the algorithm. Experimental results on the NEU-DET dataset show that the average detection accuracy is 81.1, which is about (YOLOv5) 6% higher than the original model and satisfies real-time detection. Therefore, DBCW-YOLO has good overall performance in the steel surface defect detection task.

Keywords: defect detection; steel surface; CARAFE; BiFPN; WIoU; DyHead; YOLOv5

Citation: Han, J.; Cui, G.; Li, Z.; Zhao, J. DBCW-YOLO: A Modified YOLOv5 for the Detection of Steel Surface Defects. *Appl. Sci.* **2024**, *14*, 4594. https://doi.org/10.3390/app14114594

Academic Editor: Luis Javier Garcia Villalba

Received: 7 May 2024
Revised: 23 May 2024
Accepted: 24 May 2024
Published: 27 May 2024

1. Introduction

Steel is an important raw material that plays an important role in industrial manufacturing. Therefore, ensuring steel quality is a crucial and demanding task. During the steel manufacturing process, the production environment and processing equipment limitations can result in various surface defects on the product, such as cracks, scratches, plaques, punches, indentations, and other imperfections. These defects can affect both the aesthetics and quality of steel [1]. The detection of defects on the surface of steel is, therefore, an essential part of industrial production.

The earliest method of defect detection was manual visual inspection. However, the traditional manual visual inspection method suffers from high subjectivity and empirical variation. This may limit the reliability of the test results. In addition to this, manual inspection is inefficient and costly, which limits the further development of traditional manual visual inspection methods. As machine learning continues to advance, defect detection methods on the basis of machine learning gradually replace manual detection methods. An adaptive method for detecting steel surface defects by exploiting the Haar wavelet transform was proposed by Xu et al. [2] and was fruitful. Ai et al. [3] used the features statistically derived from the magnitude spectrum obtained by Fourier transform for crack detection on the steel plate surface. In another approach, Medina et al. [4] used Gabor filters for spatial and frequency domain defect detection in steel coils. These methods have made great strides compared to manual testing. However, machine learning methods require different analytical processing for specific images, resulting in poor robustness and suboptimal detection accuracy using machine learning defect detection.

Over the past few years, deep learning has achieved considerable advances in flaw detection thanks to its powerful learning capabilities [5,6]. Deep learning-based object detection methods are mainly classified into one- and two-stage methods. A two-stage model, such as the R-CNN series [7–9], follows a two-step model, first generating candidate regions and then classifying them after refining their positions. The two-stage detection method performs well on the detection error rate and missed detection rate. However, the speed of detection is relatively slow, and it is not able to achieve the requirements of real-time detection. Therefore, one-stage detection methods have emerged. The widely used one-stage object detection models at this stage include SSD [10], YOLO series [11–14], and Retina Net [15]. One-stage target detection algorithms have gained popularity in target detection applications that require efficient and real-time performance due to their fast detection speed, end-to-end training, fewer hyperparameters, applicability to multiple tasks, excellent small target detection capability, and better real-time performance. YOLO series algorithms are more representative algorithms inside one-stage object detection, but the accuracy of one-stage target detection algorithms is insufficient.

While target detection algorithms have shown high accuracy in detecting defects with small-scale variations, they still perform poorly in detecting targets with large-scale variations. Most current target detection algorithms rely heavily on the prediction of feature mappings that provide limited information about multi-scale targets. Therefore, detailed image information needs to be utilized wisely. For example, deep networks can capture more comprehensive semantic information. However, they may be less suitable for detecting defects in ground resolution. Therefore, the performance of feature extraction networks in capturing multi-scale features is of particular interest. Moreover, for the problem of target feature loss, we can conclude that the improvement of the detection head is essential.

YOLO still fails to detect complex defects with sufficient accuracy. YOLO still needs to be optimized for improved detection accuracy. Thus, this research aims to design a steel defect detection model that can ensure high detection accuracy and a reasonable detection speed.

Based on these characteristics, to enhance the accuracy of defect detection, this paper introduces a new one-stage inspection model, DBCW-YOLO, on the basis of YOLOv5. The algorithm uses YOLOv5 as the baseline model, and for the up-sampling part, an up-sampling module (CARAFE) [16] is proposed to enhance the receptive field and obtain much semantic information. For the YOLOv5 head, add the dynamic detection head (DyHead) [17] to enhance the detection ability of the original. For the YOLOv5 model's loss function, the WIoU [18] is used to improve the baseline model's training stability, improving the model's training efficiency.

Therefore, the main contributions are listed as follows:

1. Enhanced feature fusion capability using cross-scale connectivity and embedding lightweight up-sampling (CARAFE) into the YOLOv5 network to cope with the steel defect fusion capability with a large scale of variation and to ensure the lightness of the network by improving the receptive field.
2. We use the dynamic non-monotonic focusing mechanism to replace the CIoU boundary loss function in the original model with the WioU, which enhances the competitiveness of middle-quality anchor frames and simultaneously reduces the harmful gradient generated by low-quality examples.
3. Embed the self-attention mechanism detection head (DyHead) into the YOLOv5 detection stage to enhance the detection ability of the model.

Our model is targeted to improve the characteristics of steel defects. First, BiFPN and the up-sampling module, CARAFE, are used to enhance the algorithm's focus on multi-scale information for steel surface defects with large-scale variations. Second, to address the inadequacy of the CIoU aspect ratio of the original model loss function, we introduce a WIoU to enhance the capability of the boundary loss function. For the weak detection ability of the model, we embed a dynamic detection head (DyHead) to improve the detection

ability of the model. In addition, appropriate ablation experiments are designed to validate the effectiveness of the models and the individual modules. The experimental results indicate that DBCW-YOLO can maintain high detection accuracy while also having real-time detection capability. Experiment results denote that DBCW-YOLO has an mAP of 81.1% and 33.8 FPS (frames per second) an mAP improvement of approximately 6% over the YOLOv5 model. It can provide a solution to the problem of low defect detection on steel surfaces due to large changes in defect size and strong background interference in industrial scenarios.

2. Related Work

2.1. Traditional Method

There are two main steps in machine learning-based methods. Firstly, the feature extraction rules are designed according to different types of defects for feature extraction, and then the features are inputted into the classifier to realize defect classification. A framework for extracting features of steel surface defects hidden in non-uniform patterns was proposed by Luo et al. [19]. By introducing the generalized complete local binary pattern (GCLBP), an improvement of the complete noise invariant local structure pattern (ICNLP) was made. The defect identification classification was achieved using the nearest neighbor classifier. Liu et al. [20] improved the contour transform and kernel spectral regression for metal surface defect detection by enhancing feature extraction in a multi-scale subspace. Wang et al. [21] used a guiding template to detect strip surface defects. They sorted the image by grayscale and subtracted the sorted test image from the guide template to segment strip surface defects. Inspired by the bootstrap template, the accuracy of defect detection can be improved by elevating the focus on the defective region. Cardellicchio et al. [22] proposed an adopted high-throughput data acquisition using the laser profilometry processing method and proposed a lightweight machine learning algorithm for defect detection, which is capable of high-precision real-time monitoring. Since traditional machine learning methods rely on artificially designed feature extraction principles, this causes poor generalization capability of the machine learning methods, which is easily affected by interference and noise, thus reducing the detection accuracy.

In fact, due to the many connections between traditional methods, it may be possible to use several traditional methods at the same time to jointly achieve the detection of defects. In general, conventional methods have strong limitations and require reanalyzing and designing feature extraction rules for different types of defects. For example, having dimensions that do not vary much or having sharp defect contours with low noise and high contrast under specified lighting environments. Machine learning possesses some robustness. However, artificial features have the disadvantage of weak characterization and poor adaptability. It is difficult to meet industrial needs using machine learning methods.

2.2. Deep Learning Method

With its accuracy and speed, deep learning target recognition algorithms are widely used in the industry. Deep learning-based target detection can be categorized into two-stage detection and one-stage detection. Two-stage detection methods generally have high localization and target recognition accuracies but slower detection speeds. A combination of ResNet50 and an improved Faster R-CNN algorithm was proposed by Wang et al. [23] for detecting steel surface defects. Three improvements are proposed by them to the Faster R-CNN, including enhanced feature pyramid networks (FPNs), spatial pyramid pooling (SPP), and matrix NMS algorithms, to obtain better performance. Li et al. [24] proposed a method for pre-processing tunnel surface images to improve their quality and avoid repeated detection. They also offered a multilayer feature fusion network to detect defects on the tunnel surface combined with the Faster R-CNN, achieving high detection precision. However, the speed of detection for the two-stage algorithm is significantly lower compared to the one-stage algorithm.

Conversely, one-stage methods have faster detection speeds but may have lower accuracy. Yu et al. [25] have introduced a lightweight and powerful PCB defect detection network (led-net) and built a new backbone and neck network that can efficiently fuse multi-scale features. The loss function with adaptive localization is used to calculate the localization loss and increase detection accuracy. Cheng [26] proposed a new channel attention module based on the optimized Retina Net model to enable the model to acquire more essential channel characteristics. The Adaptive Spatial Feature Fusion (ASFF) [27] module is embedded into the model, enabling the model to improve its use of spatial features. Cardellicchio et al. [28] created a bridge defect dataset and used YOLOv5 to detect bridge defects, contributing to the monitoring of bridge condition and safety. To improve the model's feature extraction ability, Li et al. [29] utilized a convolutional encoder–decoder module with residual blocks in YOLOv4 to enhance the model's feature detection ability and improve learning representation. Additionally, they designed a feature alignment module using the attention mechanism. Finally, they employed three decoupled heads for separate outputs. Lu et al. [30] used a simplified BiFPN combined with YOLO to detect citrus defects with 98.7% accuracy. Guo et al. [31] merged the TRANS module in Transformer with the YOLOv5 backbone. These features, combined with global information, improve the model's capability to dynamically adjust to objects at different scales. YOLOv5 achieves a better detection effect.

3. Method

3.1. Basic Model

Considering the computational resources and algorithm detection effect in industrial scenarios, after comparing the YOLO series of algorithms, we chose the lighter YOLOv5m 6.0 as the improved benchmark model. Its main network structure is illustrated in Figure 1. The network contains four main parts. In the input part, several key data enhancement and processing technologies are adopted. Among them, the mosaic data enhancement increases the variety and complexity of the training samples by stitching together four randomly selected images to create a single large image. Adaptive anchor frame calculation dynamically adjusts anchor frame size and position according to target size and position. For the input portion of the backbone, adaptive image sizing is used to dynamically adjust the input to satisfy the backbone section requirements. Then, after pre-processing and image enhancement operations are completed on the images, the images are input to the backbone module, which extracts features from the processed images. The neck module then fuses the acquired features, which generates three different kinds of feature information: large, medium, and small. Finally, the extracted and fused feature information is input into the head module, and the final result is output after detection.

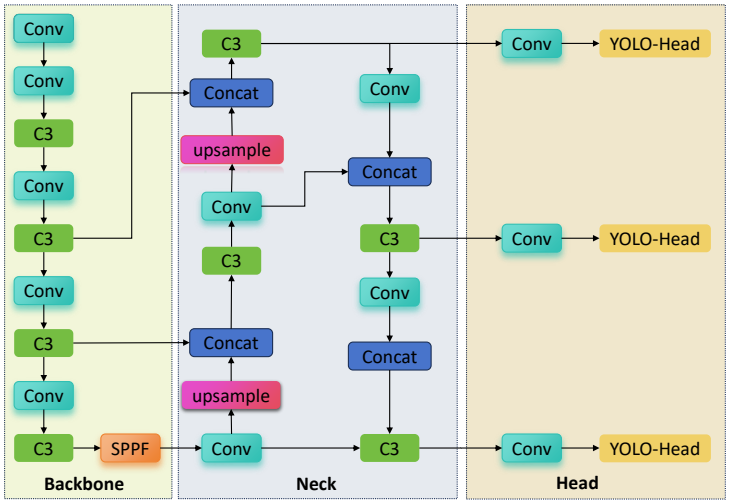

Figure 1. YOLOv5 network architecture.

3.2. DBCW-YOLO

The large-scale variation of steel surface defects and the strong background disturbance led to low discriminability of semantic information and poor detection of small targets. To enhance semantic discriminability, it is essential to obtain the scene information of the neighboring domains for information correlation and to acquire a profound comprehension of the correlation among different categories of imperfections. The DBCW-YOLO algorithm is an enhancement of the YOLOv5 algorithm, and the network structure is illustrated in Figure 2. To achieve higher detection accuracy, the model feature extraction is firstly enhanced by the strategy of BiFPN cross-scale connectivity, and the up-sampling algorithm of the YOLOv5 neck is optimized using the CARAFE module structure, which enhances the expression of features and improves the model's ability to capture contextual information. Secondly, for the large variation of sample quality and poor detection, the WIoU loss function is introduced to reduce the impact of sample quality and improve the efficiency of the model. Finally, to improve the representation of the model head, the DyHead module is introduced in the head to improve the steel defect detection performance.

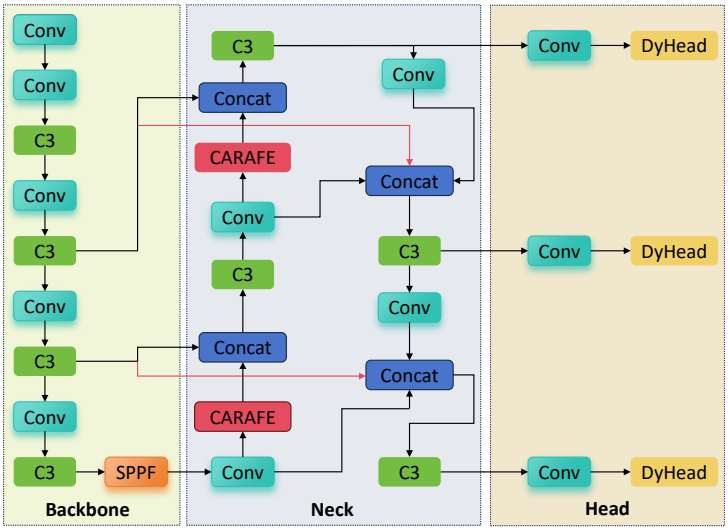

Figure 2. DBCW-YOLO network architecture.

3.2.1. Improved Feature Fusion

Feature fusion performs an essential role in target detection tasks. In YOLOv5, the PANet architecture is utilized for feature fusion. This cascade of feature maps transformed by the same size is not fully utilized for features between different sizes, making the detection accuracy limited. Moreover, the nearest neighbor interpolation method is the original up-sampling algorithm of the adopted neck network in YOLOv5. However, relying on the nearest neighbor pixel values does not allow us to obtain the subtle information of the image; the feature-aware domain is relatively small, and the edge information in the image produces an obvious jagged effect. To improve the detection abilities of the model, this paper proposes an enhanced feature fusion network, which introduces the idea of BiFPN [32] multi-scale feature fusion in the YOLOv5 neck. Moreover, this paper introduces a lightweight up-sampling module called CARAFE to improve the up-sampling algorithm for feature fusion in YOLOv5 without incurring additional computational costs.

BiFPN uses bidirectional cross-scale connectivity and weighted feature map fusion to optimize the model. Bidirectional fusion is used to construct top-down and bottom-up bidirectional channels to fuse information from different scales of the backbone network. The fusion scales are up-sampled and down-sampled for the same feature resolution scale, and horizontal connections are added between the input and output nodes of the same feature to fuse as many features as possible simply without increasing the cost. In this study, the strategy of BiFPN is used, which establishes forward and backward cross-layer feature transfer paths at different layers using bidirectional connectivity to enhance

semantic representation and differentiation. The use of shallow features and fusion of multi-scale information are improved to enhance the model's ability to recognize targets at different scales.

The structure of PANet is shown in Figure 3a. BiFPN's structure is shown in Figure 3b.

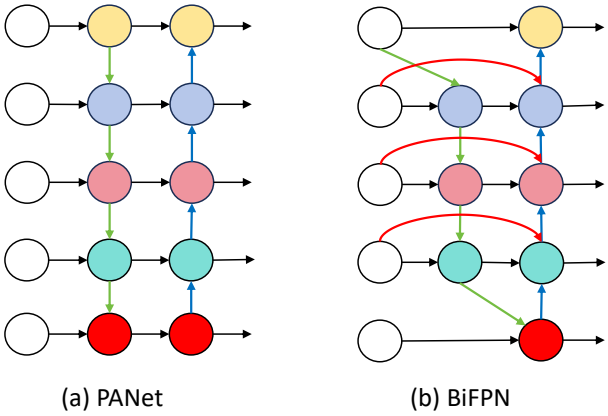

Figure 3. Comparison of structural diagrams.

For the problem of lost up-sampling information, this paper adopts a lightweight up-sampling module, CARAFE, to enhance the up-sampling algorithm of YOLOv5, which fully captures the semantic information in steel defect images and enhances the feature mapping capability, and it does not require more computational cost.

CARAFE is an up-sampling operator that utilizes feature adaptation and feature reorganization. It is mainly composed of two parts: a content-aware reorganization module and a kernel prediction module. Its function is mapped from the input features of shape $H \times W \times C$, and the feature map with shape $\delta H \times \delta W \times C$ (δ denotes the up-sampling ratio) is output by up-sampling kernel prediction and feature reorganization. Moreover, the newly generated feature map includes more semantic information. The network comparison of the original network and the improved network is illustrated in Figure 4.

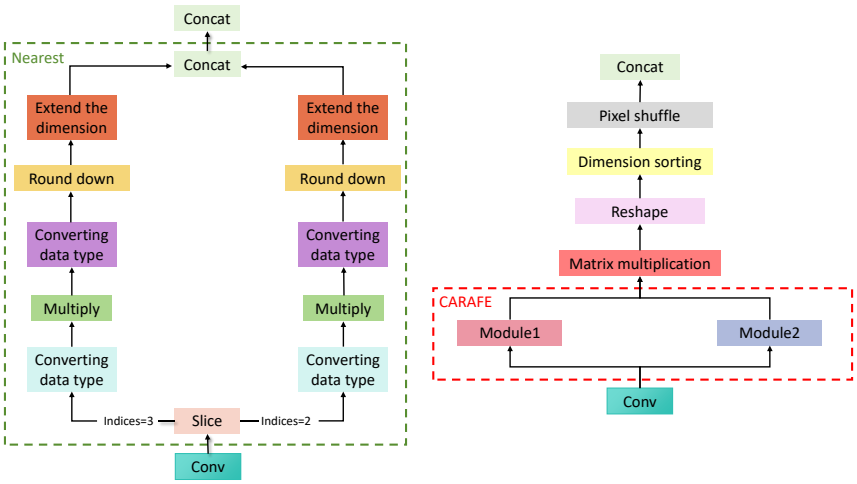

Figure 4. Original network (**left**) and improved network (**right**).

In Figure 4, Module 1 stands for the part of the kernel prediction module and Module 2 stands for the part of the content-aware reassembly, whose structures are illustrated in Figure 5. The parameters are explained as follows: N: batch size, C: input channel of the feature mapping, H: image height, W: image width, Cm: compression channel, k_{en}^2: encoder size, δ: up-sample ratio, and k_{up}^2: recombination core size.

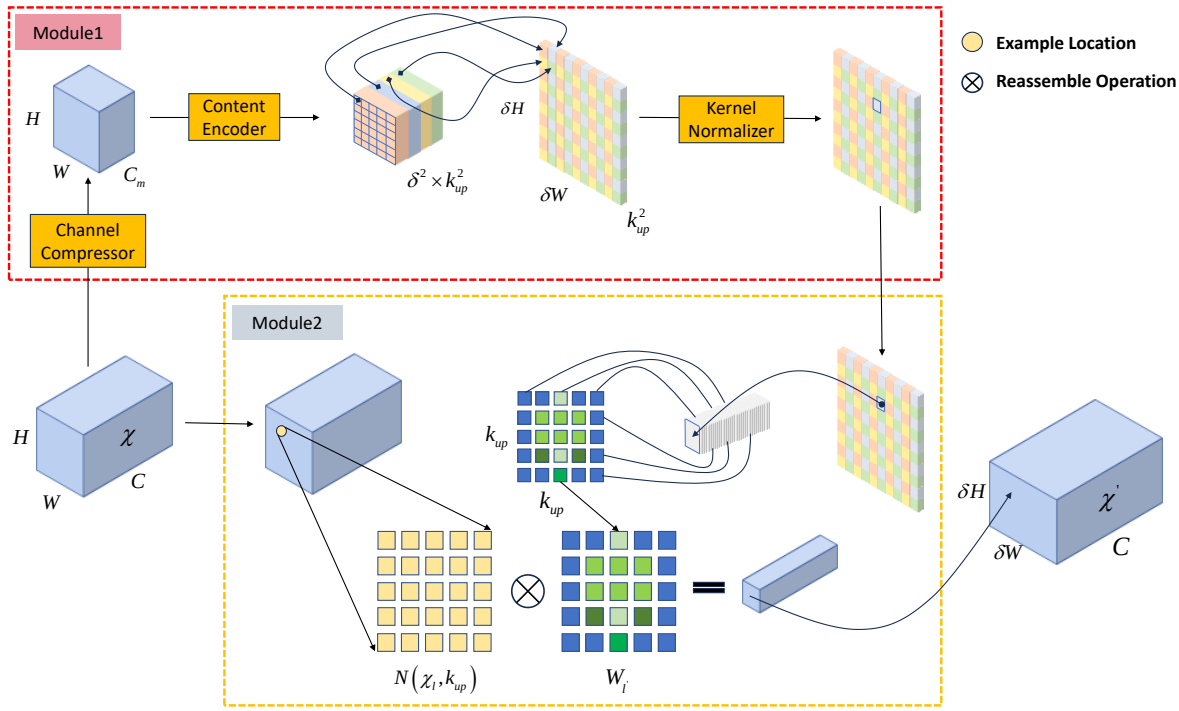

Figure 5. The structure of CARAFE.

The function of the kernel prediction module is to generate a reorganized convolutional kernel. The input feature mapping is first compressed by a 1×1 convolution operation to reduce the computational effort. Next, the compressed input feature mapping is up-sampled for kernel prediction using an encoder, and the channel dimensions are expanded in the spatial dimension to gain an up-sampled kernel of shape $\delta H \times \delta W \times k_{up} \times k_{up}$. In the end, the up-sampled kernel is normalized so that its convolution weights sum to 1.

The module for reorganizing content-aware maps each location in the output feature map back to the input feature map. Then, the region centered at $k_{up} \times k_{up}$ is taken out, and the up-sampled kernel at that point after prediction is made dot product to gain the output value. The same up-sampling kernel is used for different channels in the same position.

All calculation parameters is $2(C+1)C_m + 2(C_m k_{en}^2 + 1)\delta^2 k_{up}^2 + 2\delta^2 k_{up}^2 C$.

3.2.2. DyHead

Thanks to the large differences in the scale of the steel flaws, the network head needs to have the capability to detect steel flaws at different scales. However, the YOLOv5 model contains only three detection heads, which may cause missing detection when dealing with small target detection. At present, many researchers are increasing the detection layer to four layers from the original model to ensure that the fusion of shallower feature maps has more powerful semantic information and more accurate location information. The model improves the improvement to the sensitivity of the small target in a more comprehensive and accurate detection of steel defects and provides more reliable support for industrial inspection, etc.

In YOLOv5, the backbone network outputs a three-dimensional tensor with dimensions of horizontal $\times$ space $\times$ channel. Therefore, it improvs the integration of the variety of feature scales due to the difference in target scales and the different types and spatial positions of the object contained in the potential positional relationship features. This paper introduces the dynamic head block (DyHead) in the neck section. The DyHead enables dynamic detection of scale, space, and task awareness attention simultaneously. That is, an attention method is applied to each specific dimension of the feature tensor. The

three-dimensional feature tensor is given on the detection layer $F \in R^{L \times S \times C}$. The attention function is calculated in Equation (1) as follows:

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \tag{1}$$

where W represents the attention function, L stands for the level of the feature graph, S stands for the result of multiplying the height and width of the feature graph, and C stands for the channel numbers in the feature graph. $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ are three attention functions applied to dimensions L, S, and C. These three attention sequences are applied to the detection head and can be used multiple times in superposition. In this paper, two groups of $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ modules are superimposed successively to enhance the representation effect of the detection head and to improve the detection ability of the model for small flaws. Only two groups are added to ensure the calculation amount of the model. The single DyHead structure is shown in Figure 6.

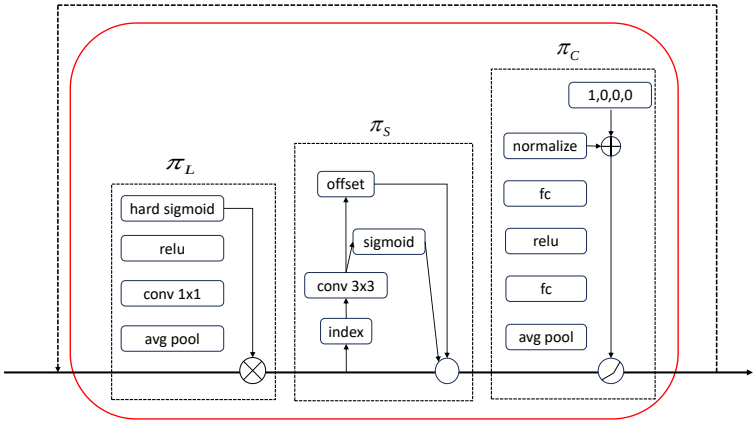

Figure 6. A specific structure of the DyHead.

The computational procedure for each of the three attention modules is as follows:

$$\pi_L(F) \cdot F = \delta\left(f\left(\frac{1}{SC}\sum_{S,C} F\right)\right) \cdot F \tag{2}$$

$$\pi_S(F) \cdot F = \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{K} W_{l,j} \cdot F(l; p_j + \Delta p_j; c) \cdot \Delta m_j \tag{3}$$

$$\pi_C(F) \cdot F = max\left(\alpha^1(F)F_C + \beta^1(F), \alpha^2(F) \cdot F_C + \beta^2(F)\right) \tag{4}$$

In Equation (2), f is a linear function composed of approximately convolution operations to achieve feature dimensionality reduction and $\delta(x)$ is the activation function, which is a hard sigmoid. In Equation (3), K stands for the sparse number of sampling locations. $p_j + \Delta p_j$ is a movable position determined by a self-learning space displacement Δp_j used to focus on some discriminative positions and Δm_j is a self-learning importance scalar at position p_j, both of which are learned from input features at the intermediate level of F. In Equation (4), the feature slice of the channel C is F_C, and $\theta(\cdot)$ is a superfunction for activation threshold control learning. Its implementation is the same as dynamic Relu, where α, β are learnable parameters through which different channels are activated differently to achieve attention operations. These three attention mechanisms are applied sequentially in the model and can be stacked together several times to form the desired DyHead block.

3.2.3. Wise-IoU

The loss function of the Bounding Box Regression (BBR) is a key part of target detection, and the quality of detection is largely up to how the loss function is designed. As an essential

part of the bounding box loss function, its accurate definition can significantly enhance the quality of the detection part. Therefore, choosing a more appropriate loss function becomes the primary task of target detection. The YOLOv5 used is the CIoU loss.

The CIoU loss function adds the calculation of aspect ratios and does not balance the dataset itself. Calibrating steel data perfectly is difficult, and low-quality samples may exist due to their specific features. Consequently, the CIoU did not have a dynamic measure of data quality during the testing of this sample. To improve the detection accuracy, a dynamic measure of the quality of the anchor box is needed. This will overcome the shortcomings of the loss function. This article optimizes the bounding box loss function using the WIoU. The WIoU BBR loss function distinguishes the quality of the anchor box using outliers, which refer to the degree of abnormality. A smaller degree of anomaly is assigned for high-quality anchor boxes and a larger degree of outlier is assigned for low-quality anchor boxes. As a result, the data contain a greater number of anchor boxes of medium quality, which enhance the main decisions and improve the overall detector's capability. The parameter diagram of Wise-IoU is shown in Figure 7.

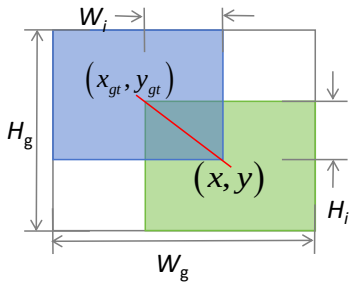

Figure 7. Wise-IoU parameter diagram.

If the anchor box can achieve a high match with the target box, then a competent loss function should mitigate the effects of geometric factors, and less intervention during model training means that the model is likely to achieve a higher generalization capacity. On this basis, the distance–attention mechanism was constructed, and a WIoUv1 with a two-layer attention mechanism was obtained.

$R_{WIOU} \in [1, e)$ will enhance the L_{IOU} of the middle-quality candidate box.

$$L_{WIOUv1} = R_{WIOU} L_{IOU} \tag{5}$$

$L_{IOU} \in [0, 1]$ overwhelmingly decreases the R_{WIOU} of the high-quality candidate box, and it focuses on the distance between the prediction box and the centroid of the candidate box when the intersection over the union (IoU) is large.

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g{}^2 + H_g{}^2)^*}\right) \tag{6}$$

where W_g and H_g are* the size of the smallest closed box (Figure 7). To prevent R_{WIOU} from creating gradients that impede convergence, W_g and H_g are separated from the computed graph (the superscript * stands for this work). Therefore, there is no need to consider the introduction of new metrics to remove barriers to convergence.

4. Experiment

4.1. Dataset

In this article, the real-world benchmark dataset, NEU-DET, is selected to complete the experiment. These data include six categories, and the number of defects in each type is 300. The six categories of defects are Crazing (CR), Inclusion (In), Patches (Pa), Rolled-in Scale (RS), and Scratches (Sc). The detection image is displayed in Figure 8.

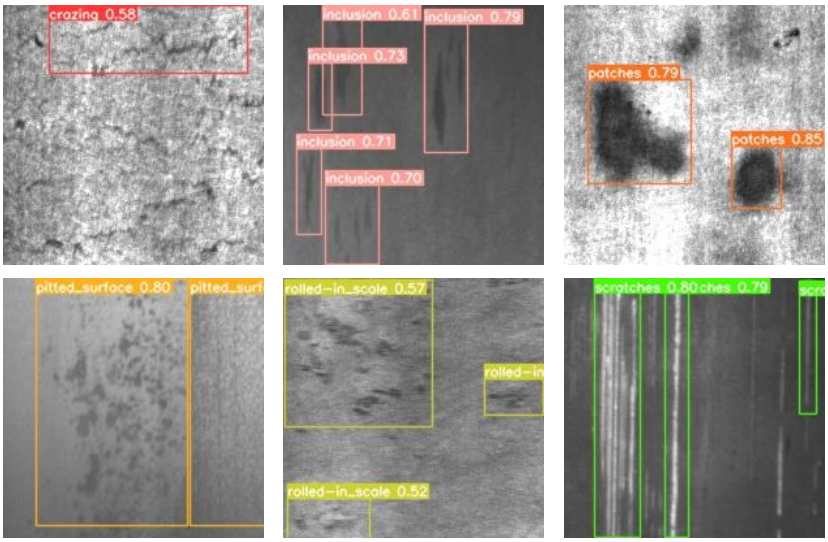

Figure 8. Detection results (DBCW-YOLO).

4.2. Index of Evaluation

To comprehensively evaluate the improvements in the algorithm's performance and to compare it with other algorithms, in this paper, several assessment indicators are used, including precision (P), recall (R), average precision (AP) for single-type precision, mean average precision (mAP) for multi-type precision, and frames per second (FPS) for detection speed. FPS is frames per second. Therefore, in this paper, experimental validation was carried out using the same equipment. Calculations of P, R, AP, and mAP are displayed in Equations (7)–(10) as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 PdR \tag{9}$$

$$L_{mAP} = \frac{1}{N}\sum_{i=1}^{N} A_P(i) \tag{10}$$

4.3. Experimental Environment

The environment and relevant parameters of the experiment are displayed in Table 1.

Table 1. Experimental environment and parameters.

Parameters	Value
Operating System	Windows 10
GPU	NVIDIA RTX 2080Ti (manufactured by NVIDIA Corporation, based in Santa Clara, CA, USA)
Framework	PyTorch 1.10.0
Optimizer	SGD
Momentum	0.937
Weight Decay	0.0005
Learning Rate	0.01
Epoch	150
Batch Size	8
Image Size	200 × 200

4.4. Experimental Result

Contrast Experiment

To prove the advantage of DBCW-YOLO, this paper uses several mainstream algorithms to compare NEU-DET datasets. In industrial applications, firstly, the detection accuracy must be guaranteed. Secondly, considering the production speed, the algorithm must have a decent detection speed. Therefore, an accuracy metric (mAP) and a FPS detection speed metric are selected to be shown in Table 2. The experimental results are displayed in Table 2.

Table 2. Detect result comparison.

Types	YOLOv3	Faster R-CNN	Retina Net	YOLOv5s	YOLOv5l	YOLOv7	YOLOv8	DBCW-YOLO
Cr	40.9%	44.7%	45.9%	42.3%	43.2%	46.3%	42.7%	**51.0%**
In	81.8%	79.2%	84.2%	79.8%	81.6%	78.1%	84.2%	**87.1%**
Pa	91.8%	82.1%	91.1%	92.4%	92.5%	88.6%	90.8%	**93.0%**
PS	**94.9%**	89.4%	88.6%	92.5%	92.9%	90.5%	89.0%	92.8%
RS	64.2%	65.3%	58.6%	54.7%	61.8%	67.7%	65.4%	**70.0%**
Sc	91.3%	89.3%	81.6%	87.2%	**96.5%**	84.6%	87.2%	92.9%
mAP 0.5	77.5%	74.6%	75.0%	74.8%	78.1%	76.0%	76.5%	**81.1%**
FPS	55.2	17.4	41.2	97.1	48	**125**	57.6	33.8

In Table 2, the best result for each detect are in bold.

From the data in Table 2, we can conclude that the algorithm in this paper has the highest accuracy in the table, reaching 81.1%. The DBCW-YOLO algorithm has the highest detection effect of four kinds of defects. Among these algorithms, YOLOv7 has the fastest detection speed, but the accuracy of each class is not very high, and the overall ability is general. The detection accuracy for all types of defects is better than the newer YOLOv8. Although YOLOv3 and YOLOv5l have good detection results in some defects, the overall average accuracy still has a certain gap compared with our proposed methods. This is because the DBCW-YOLO proposed by us can better extract features and take into account the large variation of steel defect scales. In summary, our proposed DBCW-YOLO achieves high detection accuracy and good FPS.

The result in Table 3 shows that our method surpassed the original method in most of the P and F1 in all the test items, and AP was superior in all of them, which verified the validity of our method.

Table 3. The comparison of detecting results on NEU-DET.

Methods	Type	P	R	F1-Score	AP	mAP
YOLOv5m (baseline)	Cr	56.1%	26.7%	0.362	39.8%	75.3%
	In	70.7%	89.3%	0.789	79.7%	
	Pa	81.0%	89.7%	0.851	91.6%	
	PS	85.9%	82.9%	0.844	90.5%	
	RS	51.9%	66.0%	0.581	58.6%	
	Sc	80.7%	85.5%	0.830	91.8%	
DBCW-YOLO (Improvement)	Cr	56.8%	46.3%	0.510	51.0%	81.1%
	In	71.1%	87.4%	0.784	87.1%	
	Pa	79.4%	93.0%	0.857	93.0%	
	PS	93.2%	80.5%	0.864	92.8%	
	RS	65.2%	77.4%	0.708	70.0%	
	Sc	88.9%	80.0%	0.842	92.9%	

A comparison of the DBCW-YOLO and YOLOv5 under each type is given in Figure 9. The figure illustrates the improvement of the detection results of different types of defects in the original model and the DBCW-YOLO model. The accuracy improvement of the

two types of defects, Cr and RS, which are smaller targets that are more difficult to detect, and DBCW-YOLO greatly improves the AP values of these two defects. DBCW-YOLO greatly improves the AP values of these two defects. In Figure 9, the AP of CR in the improved YOLOv5 has increased by more than 10% compared with other algorithms, and the AP of RS has also increased by 5.8% compared with YOLOv5, and the effect is powerful compared to other algorithms. This suggests that DBCW-YOLO acquires deeper features and improves results significantly for small targets. The AP values of the other four defects have good detection results compared with other algorithms. The overall mAP was 81.1 percent. In Table 2, DBCW-YOLO outperforms the other methods for most of the defects detected, and the effect is substantially improved. By comparing Figure 9, we can conclude that the overall defect detection capability of the method proposed in this paper is significantly improved, and it can meet the needs of real-time detection in the industry.

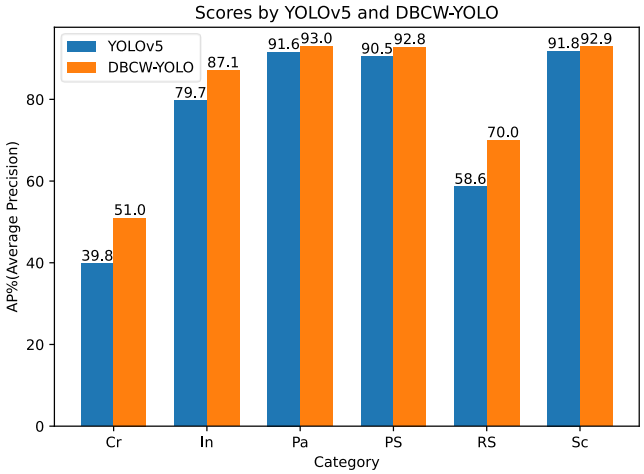

Figure 9. AP comparison of various types.

4.5. Ablation Experiment

According to Table 4, we know that our improvement is useful. The mAP value of YOLOv5m is 75.3%, and the mAP value of DBCW-YOLO is 81.1%, which has improved the effect on all six types of defects. Cr increased by 11.2%, In increased by 7.3%, Pa increased by 1.4%, PS raised by 2.3%, RS raised by 11.4%, and Sc raised by 1.1%. For the function of each module, ablation experiments were conducted in this article, respectively, and the mAP value was significantly improved by each module, while the increase in the mAP value by module superposition was still 4% and 4.3%. Therefore, our experiments proved the usefulness of every module. For the two types of defects in the detection effect of the benchmark model, Cr and Sc both increased by more than 10%. Compared to YOLOv5m, the overall mAP value of DBCW-YOLO increased by 5.8%, which verified the detection capability of DBCW-YOLO.

Table 4. Ablation experiments on NEU-DET.

Methods	mAP 0.5	Cr	In	Pa	PS	RS	Sc
YOLOv5m	75.3%	39.8%	79.7%	91.6%	90.5%	58.6%	91.8%
W-YOLO	76.8%	39.3%	80.0%	91.0%	96.0%	61.1%	93.5%
BC-YOLO	77.3%	40.6%	80.7%	95.1%	96.8%	61.7%	88.9%
D-YOLO	78.5%	47.3%	80.6%	93.4%	94.4%	61.5%	94.1%
DW-YOLO	79.3%	46.9%	88.9%	90.2%	90.5%	65.5%	93.8%
BCW-YOLO	79.6%	50.4%	82.1%	93.9%	94.5%	65.8%	91.2%
DBCW-YOLO	81.1%	51.0%	87.1%	93.0%	92.8%	70.0%	92.9%

In Table 4, W is the WIoU, BC is CARAFE and BIFPN, D is DyHead, and DW, BCW, and DBCW are their combination.

Experiments indicate that compared with the base model and other models of the network, our method improves the accuracy of defect detection in steel structures, which further proves the superiority of the DBCW-YOLO algorithm.

5. Conclusions

In this paper, the DBCW-YOLO model is presented due to the challenges of difficult image detection of small- and medium-sized defects in steel structures. In DBCW-YOLO, we propose a lightweight up-sampling method, namely, CARAFE, to enhance the baseline model. Aiming at the insufficient learning ability of the model for sample defects, a feature fusion method combining the BiFPN strategy and the lightweight up-sampling method, CARAFE, is presented. Furthermore, we introduce the WIoU to enhance the model's ability to learn weight information from feature maps. At the prediction phase, we employ a dynamic head (DyHead) to further improve the detection performance. Meanwhile, a dynamic head (DyHead) is used to improve the detection performance in the network prediction phase. Experimental results illustrate that our model achieves significant performance compared with other models.

It is worth noting that the type of steel structure selected for this experiment is relatively homogeneous, and the applicability of DBCW-YOLO could be improved. Therefore, future research will include extending the dataset to cover more different types of metal defects to improve the overall capability and adaptability of the model.

Author Contributions: J.H.: writing—reviewing and supervision; G.C.: writing—reviewing and editing, revising the manuscript, and software; Z.L.: editing and plotting the figures; J.Z.: providing valuable input and discussions in the early stages of the research. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by an intelligent monitoring and decision-making system for train operation status at the station (No. 23YFZXYC00028) and research on the design of a sample bank based on intelligent detection of railway trains (No. 17JCTPJC53500).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

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

References

1. Bozic, J.; Tabernik, D.; Skocaj, D. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Comput. Ind.* **2021**, *129*, 103459. [CrossRef]
2. Xu, K.; Xu, Y.; Zhou, P.; Wang, L. Application of RNAMlet to surface defect identification of steels. *Opt. Lasers Eng.* **2018**, *105*, 110–117. [CrossRef]
3. Ai, Y.H.; Xu, K. Surface Detection of Continuous Casting Slabs Based on Curvelet Transform and Kernel Locality Preserving Projections. *J. Iron Steel Res. Int.* **2013**, *20*, 80–86. [CrossRef]
4. Medina, R.; Gayubo, F.; González-Rodrigo, L.M.; Olmedo, D.; Gómez-García-Bermejo, J.; Zalama, E.; Perán, J.R. Automated visual classification of frequent defects in flat steel coils. *Int. J. Adv. Manuf. Technol.* **2011**, *57*, 1087–1097. [CrossRef]
5. Tulbure, A.-A.; Dulf, E.-H. A review on modern defect detection models using DCNNs—Deep convolutional neural networks. *J. Adv. Res.* **2022**, *35*, 33–48. [CrossRef]
6. Hassaballah, M.; Awad, A.I. *Deep Learning in Computer Vision: Principles and Applications*; CRC Press: Boca Raton, FL, USA, 2020.
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

9. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
12. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
13. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
14. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
15. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
16. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware Reassembly of Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2020.
17. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic Head: Unifying Object Detection Heads with Attentions. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
18. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
19. Luo, Q.W.; Sun, Y.C.; Li, P.C.; Simpson, O.; Tian, L.; He, Y.G. Generalized Completed Local Binary Patterns for Time-Efficient Steel Surface Defect Classification. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 667–679. [CrossRef]
20. Liu, X.M.; Xu, K.; Zhou, D.D.; Zhou, P. Improved contourlet transform construction and its application to surface defect recognition of metals. *Multidimens. Syst. Signal Process.* **2020**, *31*, 951–964. [CrossRef]
21. Wang, H.Y.; Zhang, J.W.; Tian, Y.; Chen, H.Y.; Sun, H.X.; Liu, K. A Simple Guidance Template-Based Defect Detection Method for Strip Steel Surfaces. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2798–2809. [CrossRef]
22. Cardellicchio, A.; Nitti, M.; Patruno, C.; Mosca, N.; di Summa, M.; Stella, E.; Renò, V. Automatic quality control of aluminum parts welds based on 3D data and artificial intelligence. *J. Intell. Manuf.* **2023**, *35*, 1629–1648. [CrossRef]
23. Wang, S.; Xia, X.J.; Ye, L.Q.; Yang, B.B. Automatic Detection and Classification of Steel Surface Defect Using Deep Convolutional Neural Networks. *Metals* **2021**, *11*, 388. [CrossRef]
24. Li, D.W.; Xie, Q.; Gong, X.X.; Yu, Z.H.; Xu, J.X.; Sun, Y.X.; Wang, J. Automatic defect detection of metro tunnel surfaces using a vision-based inspection system. *Adv. Eng. Inform.* **2021**, *47*, 101206. [CrossRef]
25. Yu, Z.L.; Wu, Y.X.; Wei, B.Q.; Ding, Z.K.; Luo, F. A lightweight and efficient model for surface tiny defect detection. *Appl. Intell.* **2023**, *53*, 6344–6353. [CrossRef]
26. Cheng, X.; Yu, J. RetinaNet with Difference Channel Attention and Adaptively Spatial Feature Fusion for Steel Surface Defect Detection. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–11. [CrossRef]
27. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
28. Cardellicchio, A.; Ruggieri, S.; Nettis, A.; Mosca, N.; Uva, G.; Renò, V. On the use of YOLOv5 for detecting common defects on existing RC bridges. In *Multimodal Sensing and Artificial Intelligence: Technologies and Applications III*; SPIE: Bellingham, WA USA, 2023; pp. 134–141.
29. Li, S.; Kong, F.; Wang, R.; Luo, T.; Shi, Z. EFD-YOLOv4: A steel surface defect detection network with encoder-decoder residual block and feature alignment module. *Measurement* **2023**, *220*, 113359. [CrossRef]
30. Lu, J.Q.; Chen, W.D.; Lan, Y.B.; Qiu, X.F.; Huang, J.W.; Luo, H.X. Design of citrus peel defect and fruit morphology detection method based on machine vision. *Comput. Electron. Agric.* **2024**, *219*, 108721. [CrossRef]
31. Guo, Z.X.; Wang, C.S.; Yang, G.; Huang, Z.Y.; Li, G. MSFT-YOLO: Improved YOLOv5 Based on Transformer for Detecting Defects of Steel Surface. *Sensors* **2022**, *22*, 3467. [CrossRef] [PubMed]
32. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. *arXiv* **2019**, arXiv:1901.01892.