computers

MDPI

Article

Improved Deep Learning Model for Workpieces of Rectangular Pipeline Surface Defect Detection

Changxing Chen * and Afizan Azman *

School of Computer Science and Engineering, Taylor's University, Subang Jaya 47500, Selangor, Malaysia
* Correspondence: chenchangxing@sd.taylors.edu.my (C.C.); afizan.azman@taylors.edu.my (A.A.)

Abstract: This study introduces a novel approach to address challenges in workpiece surface defect identification. It presents an enhanced Single Shot MultiBox Detector model, incorporating attention mechanisms and multi-feature fusion. The research methodology involves carefully curating a dataset from authentic on-site factory production, enabling the training of a model with robust real-world generalization. Leveraging the Single Shot MultiBox Detector model lead to improvements integrating channel and spatial attention mechanisms in the feature extraction network. Diverse feature extraction methods enhance the network's focus on crucial information, improving its defect detection efficacy. The proposed model achieves a significant Mean Average Precision (mAP) improvement, reaching 99.98% precision, a substantial 3% advancement over existing methodologies. Notably, the proposed model exhibits a tendency for the values of the P-R curves in object detection for each category to approach 1, which allows a better balance between the requirements of real-time detection and precision. Within the threshold range of 0.2 to 1, the model maintains a stable level of precision, consistently remaining between 0.99 and 1. In addition, the average running speed is 2 fps lower compared to other models, and the reduction in detection speed after the model improvement is kept within 1%. The experimental results indicate that the model excels in pixel-level defect identification, which is crucial for precise defect localization. Empirical experiments validate the algorithm's superior performance. This research represents a pivotal advancement in workpiece surface defect identification, combining technological innovation with practical efficacy.

Keywords: target detection; defect detection; attention mechanism; multi feature fusion

check for updates

Citation: Chen, C.; Azman, A. Improved Deep Learning Model for Workpieces of Rectangular Pipeline Surface Defect Detection. *Computers* **2024**, *13*, 30. https://doi.org/10.3390/computers13010030

Academic Editor: Leandros Maglaras

Received: 14 December 2023
Revised: 17 January 2024
Accepted: 19 January 2024
Published: 21 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the industrial field, it is difficult to avoid the defects of industrial production workpieces. On the production line, there are many kinds of workpieces, and the size is getting smaller and smaller. Often, workers with experience in workpiece detection and classification want to identify artifacts. With the continuous improvement of product quality requirements, the production of workpiece is affected by many comprehensive factors, such as the quality of the material of the workpiece and the processing equipment. Artifacts need to be classified and tested. In order to improve the efficiency of workpiece generation, ensure the detection of defective products in production engineering [1,2], and reduce the unqualified rate of workpiece products, there are different degrees and types of defects on the surface of the workpiece, which will affect the quality, safety, and performance of the workpiece [3,4]. Product quality inspection is an important part of industrial production. At first, people mainly used traditional machine learning algorithms to study this problem. With the large-scale application of deep learning models such as Convolutional Neural Networks (CNNs) in the field of computer vision [5], the use of deep learning methods for defect detection has gradually become a hot research direction. This study focuses on the surface of stainless steel rectangular pipes, hereafter referred to as "workpieces". The objective is to propose an enhanced model that combines attention mechanisms and multi-feature fusion to improve the precision of defect identification on workpiece surfaces. This approach has practical value.

2. Related Work

In order to promote the development of defect detection, researchers have put forward many effective target detection methods [6]. Some of these methods use special subnetworks to provide candidate target location suggestions to improve the detection precision of the network, and some generate higher resolution super-resolution images for small target detection through subnetworks. However, the complex subnet structure not only improves the detection performance, but also increases the number of parameters to the network, which will undoubtedly seriously reduce the speed of target detection.

Some methods based on multi-scale characterization enhance the detection capability of the network by making full use of the useful information in the network feature graph [7], which significantly reduces the operation cost. The proposed attention mechanism has introduced new vitality into the research of target detection. The attention-based method makes the network focus more attention on the target area of interest, and effectively improves the detection performance of the target detection network. The principle of the Single Shot MultiBox Detector (SSD) algorithm is to provide the classification and location information of the target directly through the backbone network, which exhibits the superior performance of simultaneously achieving high detection precision and speed [8,9]. This network structure can accommodate various sizes of targets, but there are also certain issues that need to be improved. SSD detects small targets in the underlying network, as the features in the underlying network have higher resolution compared to the high-level network, providing more specific positional spatial information for small targets. However, the feature representation of small targets in the underlying network is insufficient, and the feature information contained is not rich, causing trouble during actual testing. The research in [10] proposed to improve the structure of SSD networks, which optimizes traditional convolutional layers and effectively enhances the training effect and convergence speed of SSD networks. Nowadays, it has been widely used in many target classification and detection tasks.

Most studies have applied attention mechanisms to conventional size object detection tasks, and small object detection methods based on attention mechanisms have become an urgent research direction. Based on this, this paper proposes an improved SSD defect detection model that combines feature fusion, and adds hollow convolutional units to expand the receptive field, so that each convolutional output contains a large range of information while improving model precision.

The following Table 1 provides a performance comparison of deep learning models. Different models show differences in performance in terms of feature extraction, multi-scale object detection, and robustness in low-light conditions. References [7,10] focus on feature extraction and fusion, with accuracies of 75.3% and 76.48%, respectively. References [8,9] emphasize layered feature fusion for multi-scale object detection, achieving accuracies of 51.23% and 94.16%, respectively, indicating room for improvement. Reference [11] excels in multi-scale shallow feature fusion but requires enhanced robustness in low-light scenarios, with a precision of 80.42%. Reference [12] optimizes the model's loss function by incorporating inhibitory loss and achieves a precision of 83.50%, yet improvements are needed in prediction boxes. The analysis of these model strengths and weaknesses suggests that enhancing precision and robustness is a future research direction in the field of object detection. In addition, the choice of an appropriate model should be aligned with practical requirements and scene characteristics. The findings contribute valuable background information and analysis for the development of the target defect detection model in this study.

Table 1. Performance comparison of deep learning models.

Literature	Methods	mAP	Advantages	Disadvantages
[7]	ESA-Net	75.3%	It can extract and construct advanced features for pyramid networks.	The network can better detect small features with further improvement in performance.
[8]	FFR-SSD	51.23%	Layered feature fusion for multi-scale object detection	Precision still needs improvement.
[9]	SSD-BSP	94.16%	Integrating deep learning with computer vision	The increase in model complexity has raised computational costs.
[10]	B-FPN-SSD	76.48%	Implementing feature fusion at different scales on the feature layer	There is a slight deficiency in recognition speed.
[11]	Improved YOLOv3	80.42%	Multi-scale shallow feature fusion	The robustness in low-light conditions is not ideal.
[12]	Inception Resnet-SSD	83.50%	Incorporating inhibitory loss to optimize the model's loss function	Optimization is required for the model's prediction boxes.

3. Materials and Methods

3.1. Materials Collection

The selection of workpiece defect images mainly focuses on flat workpieces and metal material surfaces. Workpiece defect detection is the detection of various defects on the surface of a workpiece. It has the characteristics of multiple defect types, diverse defect shapes, and sizes [13]. The types of workpiece defects in this study are mainly divided into three types: inclusion, scratch, and speckle, which appear at different positions on the surface of the workpiece. The data images used in the experiment mainly come from the collection of surface defects on workpieces in the factory. Each of the three types of defects has its own characteristics: the number of speckled defects in the speckle category is not constant, the pits are small, and they exhibit a certain degree of reflectivity. The length of defects in the scratch category is inconsistent, and the direction is not certain. There are scratches of varying lengths and positions, and the depth of the scratches is very small. Under the background, the display is not obvious. As shown in Figure 1, the inclusion category has a relatively dark background due to process production reasons, making it difficult to distinguish. The width, length, and position of the inclusion defects in this category are not necessarily the same, and the color depth is also inconsistent. Therefore, the data images become very diverse during collection because there is a lot of noise in the defect background of the workpiece, including incomplete image information, low clarity, and unclear target objects. These data are not conducive to the learning and training of deep convolutional neural networks, so our laboratory conducted preliminary screening on these data, leaving 300 original images for each category.

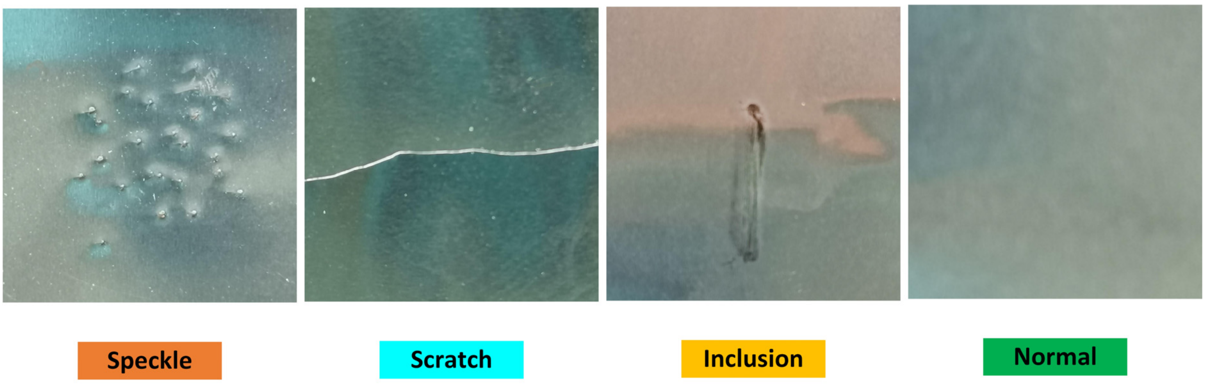


Figure 1. Defect images surface: defects and normal workpieces.

3.2. Materials Processing

The data for collecting workpiece defect images primarily come from workpieces produced by Zhaoqing No. 2 Machine Tool Factory Co., Ltd., in Zhaoqing, China. The analysis of this workpiece data aims to uncover issues that arise in practical situations, provide insights into real-world workpiece classification and defects, and obtain firsthand workpiece images. The selection of defective workpiece images mainly focuses on flat workpieces and the surfaces of metal materials. Taking into account the potential differences by binarization in the images and the need to have a sufficient number of images in the dataset to ensure that the model captures important features, data augmentation was introduced. Various transformations, including horizontal and vertical flips and color variations, were applied to the images. This process not only helps the model better adapt to different binarization levels but also enhances its generalization capabilities [14]. Therefore, after obtaining the original images, the number of images in the dataset was increased through data augmentation. Data augmentation was implemented through three primary methods: horizontal flipping, vertical flipping, and color transformation. This process generated a total of 3600 images, with 1200 images for each category. This dataset comprehensively encompasses various situations of the three types of defects in real workpieces, with strong sample representativeness. It can basically meet the requirements of diversity learning training for deep learning object detection and avoid overfitting in model training. The preprocessing of this dataset involves first performing image size transformation to adjust the image size to the input size of the model, simultaneously resizing it to maintain the aspect ratio of the original image. Otherwise, the image size needs to be directly adjusted to the input size of the model. Then, the image needs to be normalized, which scales the pixel values of the image from [0, 255] to the range of [0, 1], and performs mean removal and normalization. The specific operation is to subtract the mean of the image, and then divide it by the standard deviation to add gray bars to the image, achieving undistorted resizing.

3.3. Methods

The deep learning methods used in our study are described below.

(1) Transfer learning: The emergence of transfer learning methods is mainly due to the inability to obtain a large number of training images on their own, as deep learning relies on a large amount of annotated data. In theory, more training images are better, so in the field of image classification or object detection, a lot of time and effort is needed to annotate. At the same time, insufficient data are a common problem, due to the correlation between most data and tasks. Therefore, this correlation can be used to train new data, and transfer learning can use other large-scale data to train the obtained model weights. The learned classification parameters can be applied to another set of target domain models through transfer [15]. The transfer method used in this article is to use the weights of SSD pre-trained networks, which are the weights of the backbone feature extraction network used for feature extraction. The advantage of this is that the weights of the backbone are not random, and the feature extraction effect is more obvious. The choice of a pre-trained model as the foundation typically involves the selection of models that perform well on large-scale datasets. To adapt to the new network structure, the pre-trained model, in this case, is the SSD model [14], which was trained on the large-scale dataset. Once the pre-trained model has learned general features and patterns, the output weights are then applied to the proposed model, as shown in Figure 2.

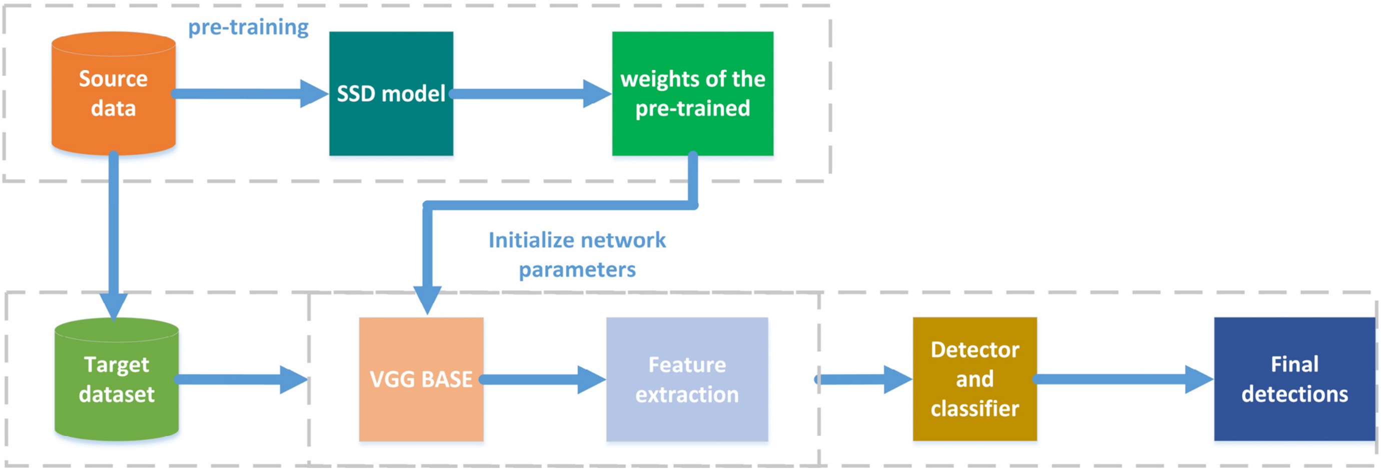


Figure 2. The process of transfer learning with a pre-trained model.

During the proposed model training process, the training is divided into two stages, namely the freezing stage and the thawing stage. The freezing stage trains parameters, and at this point, the backbone of the model is frozen, the feature extraction network does not change, unless fine-tuning the network. During the thawing phase, the parameters are trained. At this point, the backbone of the model is not frozen, and the feature extraction network will change.

(2) Feature fusion method: The feature fusion method is an important concept in deep learning networks and plays an important role in different fields of deep learning recognition tasks. The attention mechanism is an important component of feature fusion methods, which draws on human attention mechanisms and enables neural networks to selectively focus on specific parts of input data, thereby improving task execution efficiency [16]. By learning a set of weights and weighting features at different scales, the response of important features is improved. The attention mechanism allows the network to pay more attention to the target area, which can focus more on small target area information, reduce noise interference, and effectively improve the detection performance of the network [17]. The process can be represented by Equation (1).

$$\mathrm{Attention} = \mathrm{f}(\mathrm{g}(\mathrm{x}), \mathrm{x}) \tag{1}$$

This equation g() represents the attention generated by the input feature x in the focus area, while f() represents the enhancement of the input feature x in the focus area based on the attention generated by g().

The mechanism and methods used in the model are described below.

The channel attention mechanism captures specific categories of features by extracting local and global information from the image, adaptively weighting the features of different channels to enhance useful features. Its purpose is to improve the generalization ability and performance of the network by adjusting the relative weights between channels [18]. As shown in Figure 3, the channel attention mechanism is mainly composed of three parts: the Squeeze operation, the Excitation operation, and the Scale operation. The principle of the Squeeze operation is to perform global average pooling on C × H × W to obtain a feature map of 1 × 1 × C size. The next step is the Excitation operation, which performs a nonlinear transformation on the result of the Squeeze operation, using two fully connected layers to transform the number of channels. This mechanism effectively makes the model pay more attention to channel features with important information, thereby suppressing those irrelevant channel features. The last operation is the Scale operation, which can be seen as a recalibration process of the original features on the channel information. The results of the Exception operation are used as weights, and they are weighted and multiplied by the original features according to the channel.

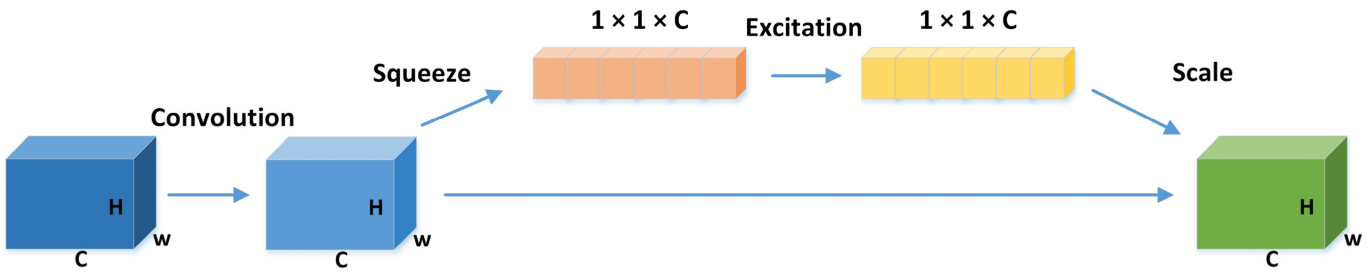


Figure 3. Squeeze-and-Excitation Attention Module.

The Convolutional Block Attention Module (CBAM) is an attention mechanism used to process channel and spatial dimension information in image or video data. The CBAM attention mechanism is shown in Figure 4, which combines the advantages of channel and spatial attention. It uses convolution operations to mix cross-channel and spatial information, and convolution operations to mix cross-channel and spatial information, and extract information features [19]. CBAM sequentially obtains effective features through the Channel Attention Module (CAM) and Spatial Attention Module (SAM). Due to the fact that CBAM consists of channel attention mechanism CAM and spatial attention mechanism SAM, effective feature extraction can be obtained from channels and positions. The CAM module focuses on the importance of each channel, and the SAM module selects meaningful local regions for each spatial location. By focusing more on specific representations, the precision of the model is improved.

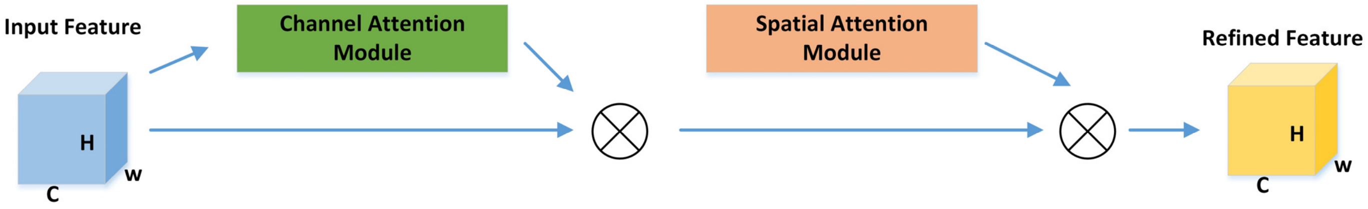


Figure 4. Convolutional Block Attention Module.

(3) ZeroPad2d zero padding method: Zero-padding2d refers to the method of zero padding the input matrix in a two-dimensional convolutional neural network to avoid pixel information loss during convolution. Its function is to effectively increase the surrounding boundary information of the image during two-dimensional convolution, so that the pixels at the boundary can receive sufficient convolution processing and avoid information loss [20].

(4) Dilated convolution: To increase the receptive field and reduce computational complexity, down-sampling is always necessary. To avoid losing resolution and still expand the receptive field, cavity convolution can be used. This is very useful in detection and segmentation tasks. On the one hand, larger receptive fields can detect and segment large targets, and on the other hand, higher resolutions can accurately locate targets [21]. Hollow convolution can arbitrarily increase the receptive field without introducing additional parameters. It is best to preserve the internal data structure and avoid down sampling, which is to retain more information while increasing the receptive field instead of pooling. The advantage is that the receptive field is increased without losing information through pooling, so that each convolution output contains a larger range of information.

3.4. Model Structure

The improved SSD model structure adopts multiple feature fusion methods to extract network feature information, and integrates multi-layer features to improve the detection performance of the network. As shown in Figure 5, the main modifications were to select channel attention and spatial main force mechanisms in the basic VGG model, and to use the hole convolution method in the fully connected layer connected to the feature extraction layer in the basic VGG model. The final fully connected layer was replaced by a global average pooling layer. In addition, the use of transfer learning can enable the model to first learn the shallow features of images on some recognized high-quality datasets, and then transfer them to its own dataset for secondary training. The network started using the pre trained weights of the entire SSD model, so it was loaded at the beginning of the algorithm training. Since this model already has good learning ability and has undergone a lot of training, it will also achieve good results when it is transferred to its own dataset for training. The improved SSD model has a much deeper network depth than the original VGG model. As the network layers continue to deepen, the feature information obtained in the images will become more complex and abstract. Therefore, theoretically, better performance can be achieved. However, due to the continuous increase in network depth, and after the modeling layers start to saturate, its precision may even decrease without increasing, which will further decrease the overall modeling performance. Therefore, six sets of channel attention and spatial main mechanism algorithm modules were added to the feature extraction layer, and two sets of zero-padding2d methods were used. Hollow convolution was also used to expand the network's perception field, which can extract complex feature information while avoiding the inability to learn just by replicating the characteristics of the previous layer network, further improving the overall performance of modeling.

The main network structure methodological approach is based on a series of carefully developed steps:

(1) The pre-trained weight input of the SSD model is used for the weight part of the backbone feature extraction network, and then the input image is used for network feature extraction.

(2) The Conv4 convolution layer uses 3 × 3 convolution processing with a 3-layer channel count of 512, and performs feature fusion processing on the 3 × 3 convolution of 512 in the second layer. After passing through the channel attention and spatial attention algorithm modules, it is then convolved again, and finally outputs 512 × 38 × 38 scale feature maps. This feature map is transferred to classification and regression processing. After passing through the pooling layer, 512 feature maps of 19 × 19 scale are output. Then, it switches to Conv5, the convolutional layer.

(3) Convolutional layer FC6 replaces the fully connected layer with a convolutional layer, allowing the network to accept input of any size. FC6 uses 3 × 3 convolutional processing with a channel count of 1024 in one layer, and then outputs 1024 × 19 × 19 scale feature maps after processing through the channel attention and spatial attention algorithm modules, which will then be transferred to the convolutional layer FC7.

(4) Convolutional layer FC7 replaces the original fully connected layer with a convolutional layer. FC7 uses 1 × 1 convolution processing with 1024 channels in one layer and dilated convolution to expand the perception field of the network. After processing, the feature map is transferred to classification and regression processing, and outputs 1024 × 19 × 19 scale feature maps, and then transitions to the convolutional layer Conv8.

(5) The Conv10 convolutional layer consists of two convolutional layers. Firstly, it undergoes 1 × 1 convolution processing with a channel count of 128, and then passes through the channel attention and spatial attention algorithm modules. After processing, it undergoes 3 × 3 convolution processing with a channel count of 256 channels and finally outputs 256 × 3 × 3 feature maps. This feature map is transferred to classification and regression processing, as well as to the Conv11 convolutional layer.

(6) The Conv11 convolutional layer is composed of two convolutional layers. First, it undergoes 1 × 1 convolution processing with a channel count of 128. Then, after passing through the channel attention and spatial attention algorithm modules, it undergoes 3 × 3 convolution processing with a channel count of 256. Finally, 256 × 1 × 1 feature maps are output, which are then transferred to classification and regression processing.

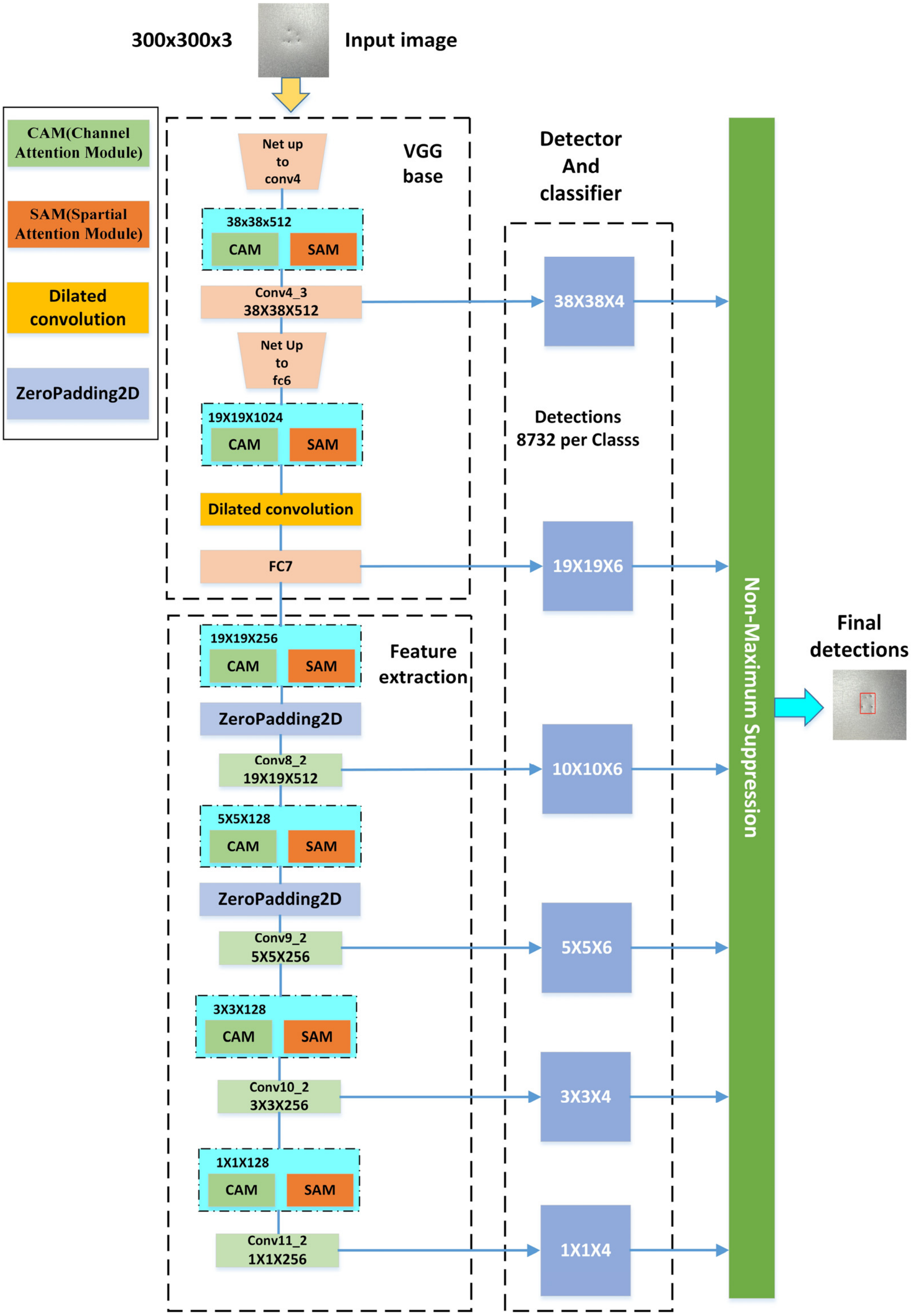


Figure 5. Improved SSD model for workpiece surface detection.

3.5. Hardware and Software Configuration

The dataset is divided into a training set, a sample validation set, and a sample testing set. According to the methods and recommendations provided in the reference literature, adjustments have been made to enhance model performance while further reducing information leakage and ensuring a more precision representation of the model's effectiveness [22,23]. The ratio of the training set to the validation set is 9:1, and the ratio of the training set to the validation set to the test set is 9:1. The input image pixels are 4032 × 3024, with epochs set to 100, the optimal number of iterations is determined through testing results, typically by monitoring the convergence of the loss value. The iterations can be concluded when the loss value stabilizes or converges, indicating that the model has achieved satisfactory performance [15,24–26]. The training is divided into two stages, namely the freezing stage and the thawing stage. The first 50 epochs are the training parameters for the freezing stage. At this time, the specific initial learning rate should be adjusted based on the particular circumstances of the task at hand. The learning rate plays a critical role in helping the model converge effectively. Experimenting with different learning rates and observing how they impact the training of the model can help determine the most suitable initial learning rate for your specific problem. It is often necessary to perform hyper-parameter tuning to find the optimal learning rate for a model [8,27,28]; the backbone of the model is frozen, and the feature extraction network remains unchanged, except for fine-tuning the network, where 0.0005 is the initial learning rate and the remaining 50 epochs are the training parameters for the thawing stage. At this point, the backbone of the model is not frozen, and the feature extraction network will change. All parameters of the network will change, and the learning rate will be set to 0.0001, where the model optimization algorithm is Adam. For some relatively small object detection tasks, setting the IOU threshold to a reasonable range, such as 0.5, to increase the sensitivity of the detection could be considered. A lower IOU threshold can make the model more likely to detect smaller or partially visible objects, but it might also result in more false positives. The choice of the IOU threshold depends on the specific requirements of the task and the trade-off between sensitivity and precision that is willing to be made [29], with 0.5 as the default IOU threshold for judging positive and negative samples. In training, the IOU threshold is set to be positive for prior boxes above 0.5 and negative for boxes below 0.5. It is used to determine whether the predicted results are correct to filter out those prediction boxes with low confidence. The network input sets the image size to 300 × 300; the specific configuration of the experimental platform is shown in Table 2.

Table 2. The configurations of the experiment.

Name	Version
CPU	Intel(R) Xeon(R) Silver 4210 CPU @ 2.20 GHz 2.19 GHz
GPU	GeForce RTX 2070 Super
Memory Bank	32 G
Operating System	Windows 10
Software environment	Cuda 10.1.1
Python Version	Python 3.7
Deep learning framework	TensorFlow2.8

3.6. Evaluation Metrics

Evaluating the performance of deep learning models involves key metrics such as true positives (TP), false positives (FP), and true negatives (TN). True positives represent instances correctly identified as positive by the model, false positives represent instances incorrectly identified as positive, and true negatives represent instances correctly identified as negative.

In practical industrial applications, the evaluation of the model involves considering its detection speed, a crucial evaluation metric expressed in frames per second (FPS). This metric serves as an indicator of the model's performance in handling data frames.

The recall refers to how many positive examples in the sample are correctly predicted. It is the proportion of correctly predicted results in all positive events. The equation is as follows. Equation (2):

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

The precision evaluation index is the proportion of the real cases in the positive cases predicted by the model, which refers to the precision. The equation is as follows. Equation (3):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

The Mean Average Precision (mAP), or the average precision of the mean, represents the mean value of the Average Precision (AP) across all defect categories. It serves as a comprehensive metric for evaluating precision. In essence, mAP is calculated by summing up the Average Precision values for all categories and subsequently dividing this sum by the total number of categories. The equation is as follows. Equation (4):

$$\text{mAP} = \frac{1}{\text{m}} \sum\nolimits_{\text{i}=1}^{\text{m}} \text{AP}_{\text{i}} \tag{4}$$

4. Results and Experimental Evaluation

4.1. Experimental Result

The training results are shown in Figure 6, where “train_Loss” represents the loss function on the training dataset, while “val_Loss” pertains to the loss function on the validation dataset. The model’s loss function exhibits a consistent decreasing trend during the initial 10 epochs. Around the 50-epoch mark, some fluctuations are observed, but the loss gradually stabilizes. By the time 100 iterations are reached, the loss function has effectively converged, displaying minimal fluctuations. This convergence signifies a successful fit of the model with a desirable level of precision.

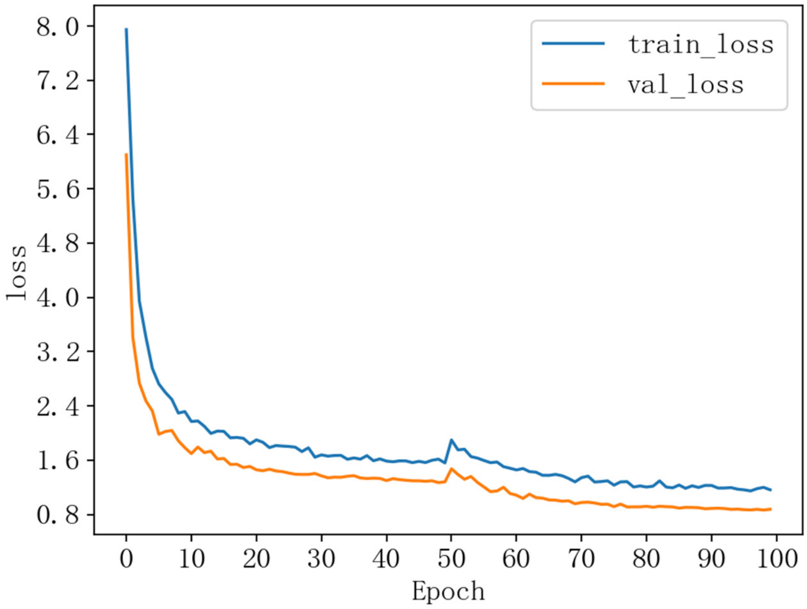
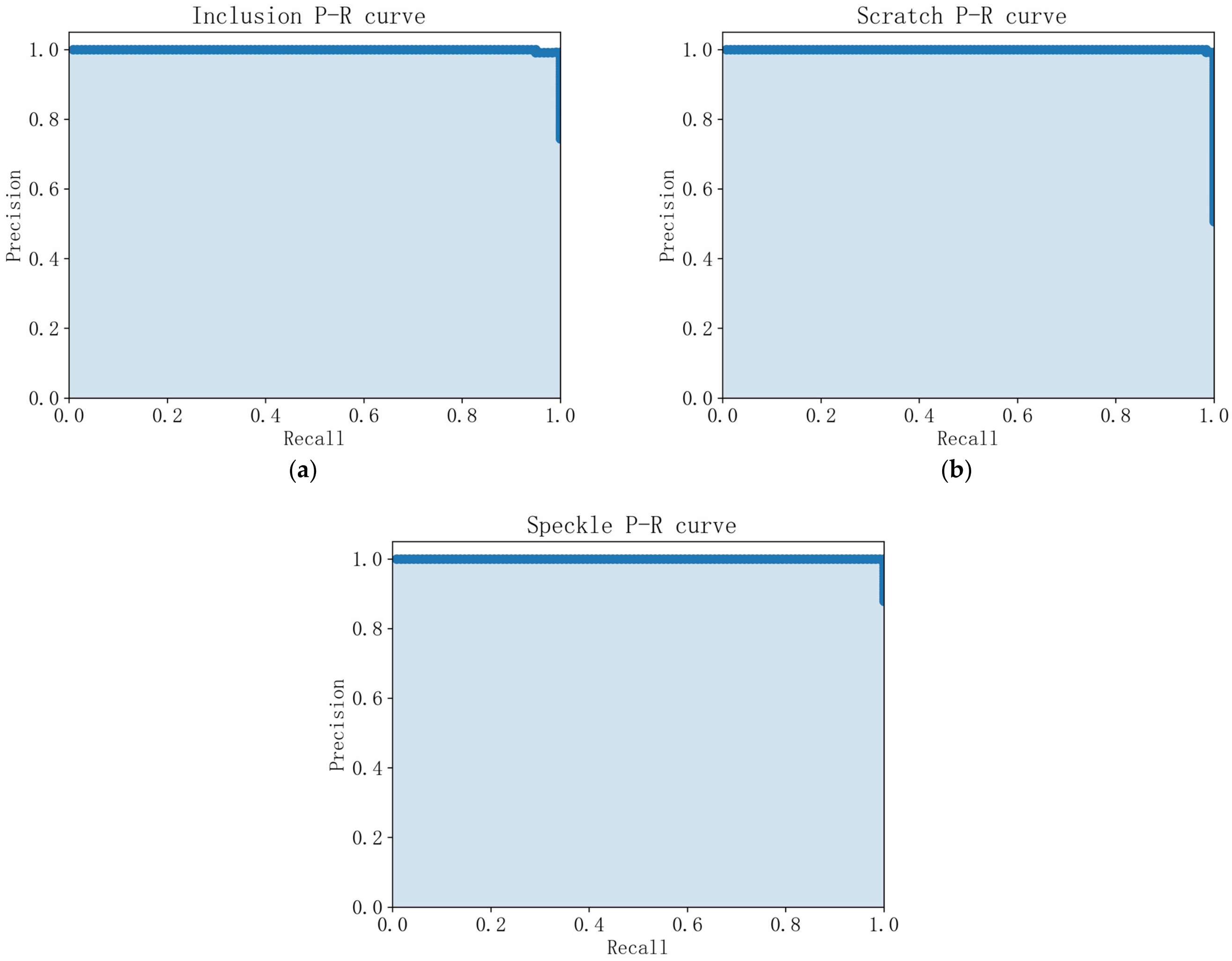


Figure 6. Loss in workpiece surface detection.

To provide a more intuitive representation of the performance of the improved SSD model in terms of precision and recall, experiments were conducted on the test set using the improved SSD model, resulting in the generation of P-R curves for detecting different defect categories. The PR curve graph, with “P” representing precision and “R” representing recall, illustrates the relationship between precision and recall. Following this pattern, the x-axis was set as recall, and the y-axis as precision. The size of the area enclosed by this line graph and the x- and y-axes is the 0–1 value, where a larger area indicates higher precision for the corresponding label type, and higher precision implies better model performance [30]. As shown in Figure 7, the experimental results show that the AP values of the P-R curves for each category tend to approach 1, and the detection precision for each category exceeds 99%. This indicates that the improved SSD model is highly effective in enhancing detection performance.



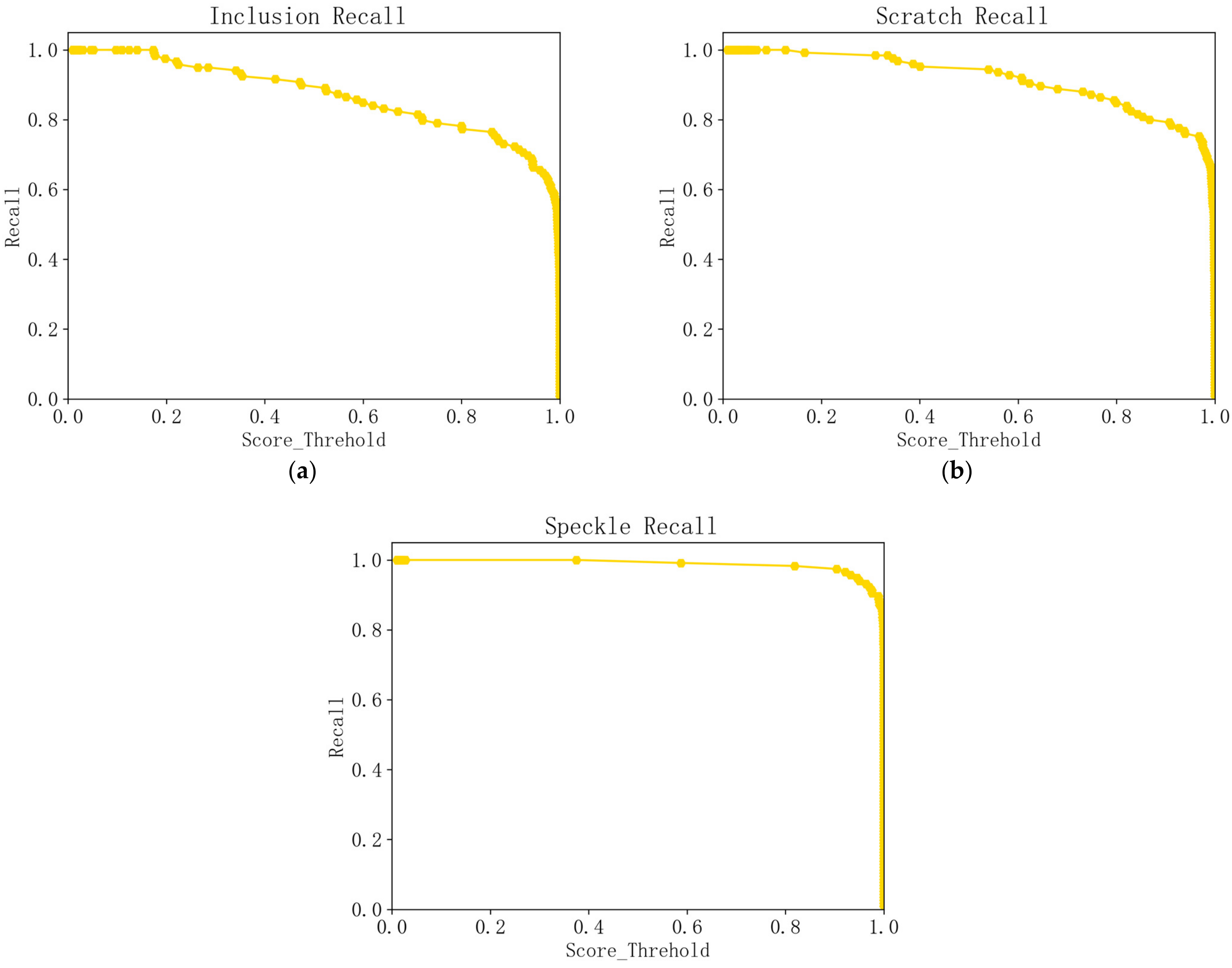
(**a**)

(**b**)

(**c**)

Figure 7. (**a**) Inclusion P-R curve; (**b**) scratch P-R curve; (**c**) speckle P-R curve.

The results in Figure 8 show that when the threshold is set between 0 and 0.2, the recall for all three categories is almost unaffected by the threshold. Moreover, when the threshold is in the range from 0.2 to 0.8, the “Speckle” category maintains a consistent recall rate in this range, hovering around 0.95 to 1, and is hardly affected by the threshold. For the other two categories, there is a noticeable gradient descent trend in the model’s recall. It is worth noting that when the threshold is set between 0.2 to 0.8, the recall for the “Inclusion” and “Scratch” categories are particularly affected by the threshold, showing a gradual decline. In other words, within this range, as the threshold increases, the recall continues to decrease. For the other two categories, there is a noticeable gradient descent trend in the model’s recall. It is worth noting that when the threshold is set between 0.2 to 0.8, the recall for the “Inclusion” and “Scratch” categories is particularly affected by the threshold and shows a gradual decline. In other words, within this range, the recall continues to decrease as the threshold increases.



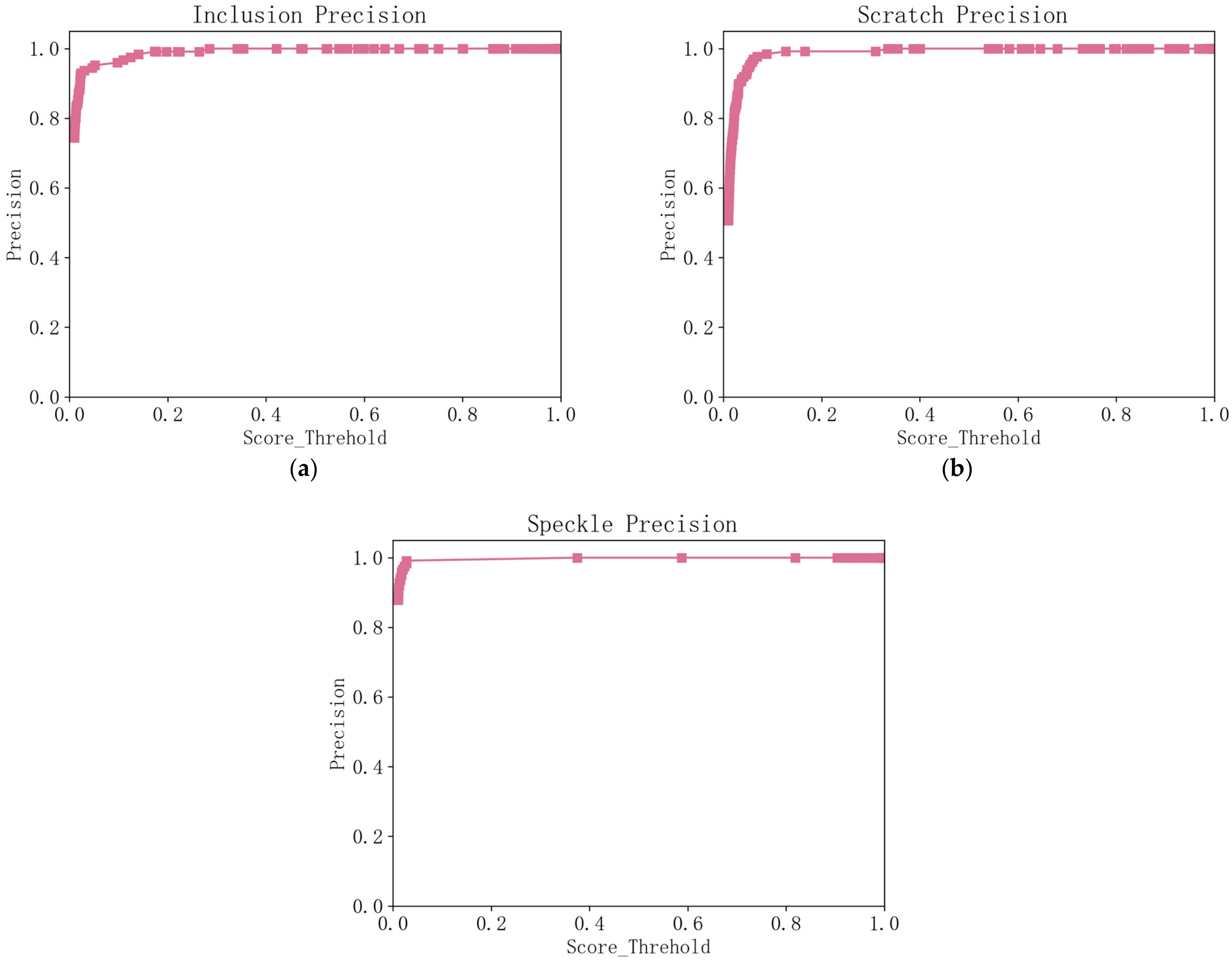
(**a**)

(**b**)

(**c**)

Figure 8. (**a**) Inclusion recall in surface detection; (**b**) scratch recall in surface detection; (**c**) speckle recall in surface detection.

Figure 9 illustrates that as the threshold increases, the improved SSD model consistently improves in precision. The most significant increase in precision occurs within the threshold range of 0 to 0.2, followed by a more gradual, steady increase. Furthermore, the model maintains a stable precision level within the range from 0.2 to 1, consistently maintaining values between 0.99 and 1. This suggests that within this range, the model's performance is relatively stable, and the threshold has a less pronounced impact on the model's precision. At a threshold of 0.2, the precision has already reached its peak and is not increasing. This indicates that a threshold of 0.2 is the highest effective threshold for the model training. In summary, within the threshold range of 0 to 0.2, the threshold has a significant effect on precision. However, in the range from 0.2 to 1.0, the threshold's influence on precision becomes progressively smaller and is almost negligible. Therefore, it is evident that a threshold of 0.2 serves as a critical threshold that marks a turning point in the model's behavior. When the threshold is set to 0.5, the model's precision for target detection is as follows: "Inclusion" is at 99.96%, "Scratch" is at 99.99%, and "Speckle" is at 100.00%. This indicates that the model has a very low rate of false positives across all categories. The model's Mean Average Precision (mAP) is 99.98%, highlighting its outstanding detection performance.



(**a**) (**b**) (**c**)

Figure 9. (**a**) Inclusion precision in surface detection; (**b**) scratch precision in surface detection; (**c**) speckle precision in surface detection.

The experimental results indicate that the improved SSD algorithm has increased the average recognition precision, as shown in Figure 10. It has improved the detection capabilities for various detection targets. The mAP has reached 99.98%, which indicates that the enhanced SSD algorithm network model can achieve a real-time and accurate detection of workpiece defects.

To validate the effectiveness of the improved SSD model, this section conducts defect detection experiments using the model. Figure 11 shows some of the predicted results for specific images using the proposed model. In complex scenarios, the proposed SSD model is able to identify defects effectively.

4.2. Error in Erea and Position

For each type of randomly selected defect for image detection (five defects), as shown in Table 3 (with the upper left corner of the image as the origin), the unit is pixel; analyzing the data in the table, the defect center location of the maximum error is in the horizontal coordinates of 20.5, the maximum error is in the vertical coordinates of 38.5, the defects are in the area of the maximum error of 4.9%. For all data, the average error of the horizontal and vertical coordinates of the defect center position is less than 2%, and the average error of the area identification is less than 5%; in the actual detection of the scope of the permissible error range, the real-time detection of defects to locate the location of positioning for the next step can be provided.

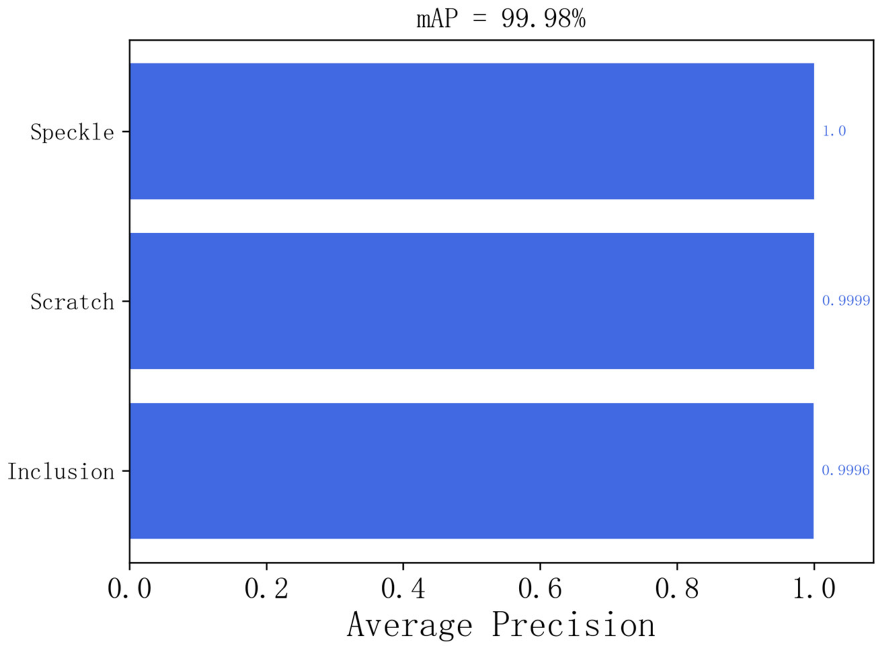


Figure 10. mAP in workpiece surface detection.

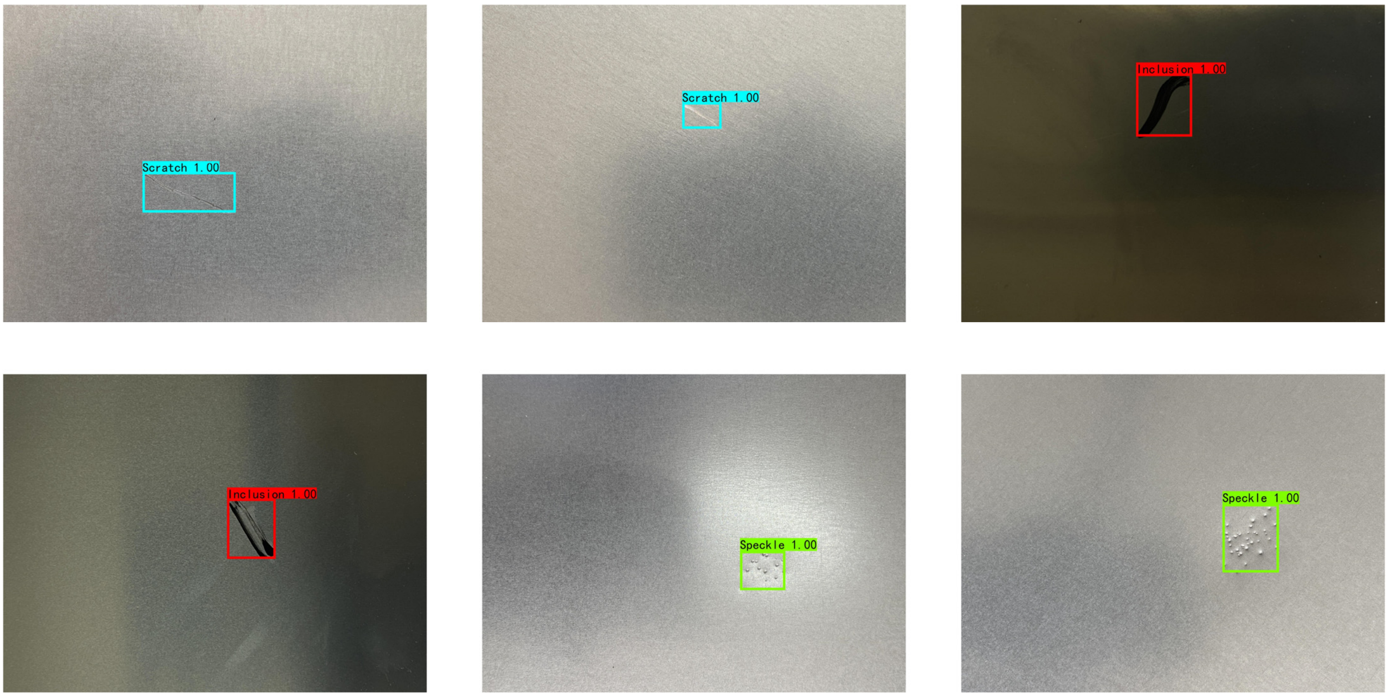


Figure 11. Prediction results.

Table 3. Error in area and position.

No.	Category	Actual Location	Identified Location	Actual Area	Identified Area	Location Error	Area Error
1	Inclusion	(2523.5, 1428.5)	(2526, 1429)	11,975	119,328	(2.5, 0.5)	147
2	Inclusion	(2790, 1328.5)	(2788.5, 1322.5)	145,350	146,157	(1.5, 6)	807
3	Inclusion	(2398.5, 208)	(2394.5, 209)	35,750	36,100	(4, 1)	350
4	Inclusion	(3004, 1166)	(2992, 1168)	194,636	197,408	(12, 2)	2772
5	Inclusion	(1932.5, 964.5)	(1930.5, 951)	326,781	319,800	(2, 13.5)	6981
6	Scratch	(2421, 1967.5)	(2441.5, 1944.5)	460,750	470,557	(20.5, 23)	9807
7	Scratch	(2624.5, 1320.5)	(2612, 1305)	387,481	401,544	(12.5, 15.5)	14,063
8	Scratch	(1732.5, 941)	(1730.5, 936.5)	349,160	335,111	(2, 4.5)	14,049
9	Scratch	(1143.5, 2316)	(1152.5, 2277.5)	135,992	142,107	(9, 38.5)	6115
10	Scratch	(2093.5, 1366)	(2086, 1374)	98,010	102,816	(7.5, 8)	4806
11	Speckle	(2754.5, 1583)	(2755, 1564)	361,020	353,760	(0.5, 19)	7260
12	Speckle	(2098, 1819)	(2090, 1815)	63,036	64,480	(8, 4)	1444
13	Speckle	(2509, 1975.5)	(2490, 1963.5)	406,334	415,950	(19, 12)	9616
14	Speckle	(2066.5, 1119)	(2070.5, 1105)	324,450	333,064	(4, 14)	8614
15	Speckle	(1787, 1914.5)	(1792.5, 1911.5)	88,796	86,355	(5.5, 3)	2441

4.3. Comparison with Other Models

To validate the detection performance of the proposed model, comparative experiments were conducted between the improved model and the original SSD model, YOLOV3, YOLOV4, and Faster R-CNN on the same dataset. The compared models use the same dataset as this study, with input image pixels set at 4032 × 3024. The initial learning rate was configured as 0.0005, and the batch size was set to 12, using the Adam optimizer. The number of prior boxes for the model was set to default, and iterations continued until the model converged before stopping. The test results, as shown in Table 4, reveal that the original SSD model achieved precision of 95.28%, while the improved network model reached a precision of 99.98%. The performance improvement is significant, and compared with the other four object detection algorithms, the improved SSD model has the highest mAP value, indicating superior detection performance. At the same time, the runtime for detection is quite satisfactory, with little difference compared to existing models, especially when compared to the time before the enhancement. Despite the increase in model complexity, the runtime differs by only 1 FPS. This indicates that the proposed SSD model has achieved excellent precision and performance with the improvements applied.

Table 4. Comparison with other models in precision.

Model	Inclusion AP/%	Scratch AP/%	Speckle AP/%	mAP%	Time (Fps)
SSD	94.56%	94.56%	97.68%	96.29%	13.27
YOLOV3	91.84%	92.32%	96.48%	93.55%	12.43
YOLOV4	92.54%	93.24%	91.34%	92.37%	19.56
Faster R-CNN	93.42%	90.36%	93.46%	92.41%	14.36
Proposed model	99.96%	99.99%	100.00%	99.98%	12.75

5. Conclusions

The recognition of defects on the surface of workpieces has problems such as difficulty in manual recognition and small targets. If the recognition is solely based on the human eye, it is prone to errors and the labor cost is particularly high, so artificial intelligence technology is needed. To address the issue of low-defect detection precision, a defect detection model based on attention mechanism and multi-feature fusion is proposed in the text. This method adds channel attention and spatial main force mechanism algorithm modules to the feature extraction network, and adopts two sets of zero-padding2d methods. It also uses hole convolution to expand the network's perception field. Through the hole convolution operation, a feature enhancement module is formed to obtain more detailed information of small-scale fault target features, while increasing the low-level feature layer perception field. In addition to enhancing the feature extraction ability of low-level feature layers for small-scale faults, it can also prevent overfitting phenomena and improve the detection performance of SSD-improved algorithms for small-scale faults, improving the situation of missed and falsely detected small defect targets. The experimental results indicate that, in terms of runtime, the average for the existing models is 14.9 FPS, which is 2 FPS higher than the proposed model. The proposed model does not show a significant decrease in speed compared to the original SSD algorithm. The decrease in detection speed is controlled within 1%. The mAP values for YOLOv3, YOLOv4, and Faster R-CNN are 93.55%, 92.37%, and 92.4%, respectively. In terms of mAP, the proposed model outperforms the existing models by at least 6%. Simultaneously, it indicates that the proposed model can accurately identify the defective targets on the surface of the workpiece, and the mAP on the defective dataset is 99.98%, which is significantly improved compared with other methods. Overall, the proposed model can better balance the requirements of real-time detection and precision. The localization position of the image is also provided, with a maximum error of 4.9% in the defect area. The average error of the horizontal and vertical coordinates of the defect center position is less than 2%. The results indicate that it can meet the requirements for position detection.

After experiments to prove that the improved target detection algorithm can be effective, a good condition for the next step is for the localization of work. Based on the theory, the method proposed in this paper can also be applied to other forms of defect detection in industrial scenes, for its performance deficiencies, which can be improved by improving these parameters which should be able to make the model more robust and generalized to adapt to more difficult detection tasks.

The application of deep learning object detection models faces certain limitations across different domains. This is due to the distinct characteristics and challenges present with defect detection problems in different domains. The model parameters proposed in this study are specifically designed for stainless steel rectangular pipes in an actual manufactured product, thus having certain scope limitations. If the model parameters are not appropriately adjusted or if the adjustment methods are inadequate, overfitting issues may occur, leading to poor performance on surface defect detection datasets with different materials. This phenomenon is quite common in the field of deep learning. Therefore, it is necessary to conduct model experiments and optimizations on specific datasets to ensure the accuracy of the model on the corresponding dataset. The proposed work not only provides a robust solution for target detection, but also lays the foundation for further refinement and innovation in the field.

Author Contributions: Conceptualization, C.C. and A.A.; methodology, C.C.; software, C.C.; validation, C.C.; formal analysis, C.C.; investigation, C.C.; resources, C.C.; data curation, C.C.; writing—original draft preparation, C.C.; writing—review and editing, C.C. and A.A.; visualization, C.C. and A.A.; supervision, A.A.; project administration, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This publication paper is funded by Taylor's University.

Data Availability Statement: The data presented in the study are included in the article.

Acknowledgments: The authors are grateful to ChatGPT, a powerful artificial intelligence tool that has provided valuable assistance in polishing our article.

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

References

1. Truong, V.D.; Xia, J.; Jeong, Y.; Yoon, J. An automatic machine vision-based algorithm for inspection of hardwood flooring defects during manufacturing. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106268. [CrossRef]
2. Yang, J.; Wang, K.; Luan, F.; Yin, Y.; Zhang, H. PreCaCycleGAN: Perceptual Capsule Cyclic Generative Adversarial Network for Industrial Defective Sample Augmentation. *Electronics* **2023**, *12*, 3475. [CrossRef]
3. Behrouzi, S.; Dix, M.; Karampanah, F.; Ates, O.; Sasidharan, N.; Chandna, S.; Vu, B. Improving Visual Defect Detection and Localization in Industrial Thermal Images Using Autoencoders. *J. Imaging* **2023**, *9*, 137. [CrossRef] [PubMed]
4. Verma, L.; Kremer, F.; Chevalier-Jabet, K. Defective PWR fuel rods detection and characterization using an Artificial Neural Network. *Prog. Nucl. Energy* **2023**, *160*, 104686. [CrossRef]
5. Mahaur, B.; Mishra, K.K.; Kumar, A. An improved lightweight small object detection framework applied to real-time autonomous driving. *Expert Syst. Appl.* **2023**, *234*, 121036. [CrossRef]
6. Zhao, Y.; Sun, X.; Yang, J. Automatic recognition of surface defects of hot rolled strip steel based on deep parallel attention convolution neural network. *Mater. Lett.* **2023**, *353*, 135313. [CrossRef]
7. Dong, S.; Teng, Y.; Jiao, L.; Du, J.; Liu, K.; Wang, R. ESA-Net: An efficient scale-aware network for small crop pest detection. *Expert Syst. Appl.* **2024**, *236*, 121308. [CrossRef]
8. Cheng, X.; Wang, Z.; Song, C.; Yu, Z. FFR-SSD: Feature fusion and reconstruction single shot detector for multi-scale object detection. *Signal Image Video Process.* **2023**, *17*, 3145–3153. [CrossRef]
9. Wang, L.; Wang, X.; Li, B. Data-driven model SSD-BSP for multi-target coal-gangue detection. *Meas. J. Int. Meas. Confed.* **2023**, *219*, 113244. [CrossRef]
10. Liu, Q.; Bi, J.; Zhang, J.; Bu, X.; Hanajima, N. B-FPN SSD: An SSD algorithm based on a bidirectional feature fusion pyramid. *Vis. Comput.* **2023**, *39*, 6265–6277. [CrossRef]
11. Cong, P.; Lv, K.; Feng, H.; Zhou, J. Improved YOLOv3 Model for Workpiece Stud Leakage Detection. *Electronics* **2022**, *11*, 3430. [CrossRef]
12. Zheng, Y. Pipeline Multitype Artifact Recognition Method Based on Inception_Resnet _V2 Structure Improving SSD Network. *Adv. Multimed.* **2022**, *2022*, 6049013. [CrossRef]
13. Xu, S.; Shao, Z. An Improved Faster RCNN based on Swin Transformer for Surface Defect Detection of Metal Workpieces. ACM International Conference Proceeding Series; Association for Computing Machinery: New York, NY, USA, 2022; pp. 120–125.
14. Tang, L.; Zhao, H.; Wang, N.; Han, B.; Wang, Y.; Li, W. An augmentation method of defect detection dataset based on M-DCGAN. *Dalian Haishi Daxue Xuebao J. Dalian Marit. Univ.* **2023**, *49*, 148–160.
15. Barbero-Aparicio, J.A.; Olivares-Gil, A.; Rodríguez, J.J.; García-Osorio, C.; Díez-Pastor, J.F. Addressing data scarcity in protein fitness landscape analysis: A study on semi-supervised and deep transfer learning techniques. *Inf. Fusion* **2024**, *102*, 102035. [CrossRef]
16. Khan, H.; Hussain, T.; Ullah Khan, S.; Ahmad Khan, Z.; Baik, S.W. Deep multi-scale pyramidal features network for supervised video summarization. *Expert Syst. Appl.* **2024**, *237*, 121288. [CrossRef]
17. Chen, S.; Lang, B.; Liu, H.; Chen, Y.; Song, Y. Android malware detection method based on graph attention networks and deep fusion of multimodal features. *Expert Syst. Appl.* **2024**, *237*, 121617. [CrossRef]
18. Hu, F.; Zhang, S.; Lin, X.; Wu, L.; Liao, N.; Song, Y. Network traffic classification model based on attention mechanism and spatiotemporal features. *Eurasip J. Inf. Secur.* **2023**, *2023*, 6. [CrossRef]
19. Zhao, G.; Zou, S.; Wu, H. Improved Algorithm for Face Mask Detection Based on YOLO-v4. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 104. [CrossRef]
20. Jia, X.; Bartlett, J.; Chen, W.; Song, S.; Zhang, T.; Cheng, X.; Lu, W.; Qiu, Z.; Duan, J. Fourier-Net: Fast Image Registration with Band-Limited Deformation. In Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, 7–14 February 2023; pp. 1015–1023.
21. Wang, X.; Wang, C.; Zhang, F.; Jiang, S.; Sun, Z.; Zhang, H.; Duan, Z.; Liu, Z. Feature fusion-based fiber-optic distributed acoustic sensing signal identification method. *Meas. Sci. Technol.* **2023**, *34*, 125141. [CrossRef]
22. Yi, C.; Liu, J.; Huang, T.; Xiao, H.; Guan, H. An efficient method of pavement distress detection based on improved YOLOv7. *Meas. Sci. Technol.* **2023**, *34*, 115402. [CrossRef]
23. Pratibha, K.; Mishra, M.; Ramana, G.V.; Lourenço, P.B. Deep Learning-Based YOLO Network Model for Detecting Surface Cracks During Structural Health Monitoring. In *RILEM Bookseries*; Springer: Berlin/Heidelberg, Germany, 2024; Volume 47, pp. 179–187.
24. Tang, J.; Wang, Z.; Zhang, H.; Li, H.; Wu, P.; Zeng, N. A lightweight surface defect detection framework combined with dual-domain attention mechanism. *Expert Syst. Appl.* **2024**, *238*, 121726. [CrossRef]
25. Shen, D.; Liu, X.; Shang, Y.; Tang, X. Deep Learning-Based Automatic Defect Detection Method for Sewer Pipelines. *Sustainability* **2023**, *15*, 9164. [CrossRef]
26. Xiao, M.; Yang, B.; Wang, S.; Mo, F.; He, Y.; Gao, Y. GRA-Net: Global receptive attention network for surface defect detection. *Knowl. Based Syst.* **2023**, *280*, 111066. [CrossRef]
27. Shen, J.; Li, G.; Kumar, R.; Singh, R. CAD Fabric Model Defect Detection Based on Improved Yolov5 Based on Self-Attention Mechanism. *Comput. Aided Des. Appl.* **2024**, *21*, 63–71. [CrossRef]
28. Sahoo, A.K.; Behera, S.; Maurya, S.; Kale, P. Detection of Physical Impairments on Solar Panel Using YOLOv5. In *Lecture Notes in Electrical Engineering*; Springer Nature: Singapore, 2023; pp. 1–12.

29. Gudhe, N.R.; Kosma, V.M.; Behravan, H.; Mannermaa, A. Nuclei instance segmentation from histopathology images using Bayesian dropout based deep learning. *BMC Med. Imaging* **2023**, *23*, 162. [CrossRef]
30. Afshar, S.; Braun, P.R.; Han, S.; Lin, Y. A multimodal deep learning model to infer cell-type-specific functional gene networks. *BMC Bioinform.* **2023**, *24*, 47. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.