# Industrial Surface Defect Detection Algorithm Enhanced with Attention Mechanism

**Jianwei Yang 23020231154243**
**Yujun Yang 23020231154246**
**XiangTao Du 23020231154180**
School of information, Xiamen University
422 Siming South Road, Siming District
Xia Men, China 361000

## Abstract

Industrial surface defect detection is a crucial component in the industrial production process. It plays a positive role in enhancing the quality of industrial products, reducing raw material wastage, and improving production efficiency. With the vigorous development of deep learning, deep learning-based object detection algorithms have taken the lead in industrial surface defect detection. Emerging deep learning object detection technologies mainly fall into two mainstream categories: single-stage and two-stage algorithms. While two-stage object detection offers higher precision, it requires longer inference times and may not meet the real-time requirements of industrial production. As object detection algorithms continue to evolve, current single-stage object detection algorithms have surpassed two-stage algorithms in both detection speed and accuracy. The outstanding representative of single-stage object detection algorithms, the YOLO series, has garnered significant attention and widespread use in both the academic and industrial sectors. Therefore, we will propose some improvements based on the YOLOv5 model for industrial defect detection:

1. Add an attention mechanism to the neck feature fusion network of YOLOv5s. The output three-scale feature maps then go through four plug-and-play attention mechanism modules, namely SE, ECA, CBAM and CA, so that the multi-scale feature maps are sent to the detection head after being weighted by attention, which improves the model's accuracy. Characteristic perception
2. Replace the original CIoU with a more appropriate EIoU as the Bounding Box loss.

## Introduction

Vision-based industrial defect detection aims to identify visible defects in various industrial products, including textiles, chips, pharmaceuticals, and construction materials(Zhang, Ding, and Yan 2011). These defects, though often small, can significantly impair the normal functionality of the products. They can occur at any stage of the industrial product's lifecycleLee et al. (2019)

We can classify industrial defects into surface defects and structural defects. Surface defects primarily occur at localized positions on the product's surface, often manifesting as texture variations, irregular regions, non-uniform patterns, or incorrect patterns(Wei, Song, and Zhang 2020) For example, surface cracks, color blocks, sparse weaving in fabrics, and printing errors in brand text. These defects can be analogized as outliers or cluster anomalies based on the pixel values' difference from the surrounding background.(Li et al. 2018) Outlier-type defects typically have distinct differences in pixel values compared to a normal image, while cluster anomalies have pixel values within a similar range to the surrounding normal areas, making them more challenging to detect. Structural defects are mainly caused by overall structural errors in the product, including deformation, misalignment, missing parts, and contamination. For example, bent wires, edge defects in diodes, or components placed in the wrong positions.

We primarily focus on research in industrial surface defect detection, which has long been one of the most important areas of study in the field of industrial vision. In recent years, with the widespread adoption of deep learning in computer vision tasks, deep learning-based methods for industrial defect detection have rapidly advanced and become mainstream. The most common detection algorithm is based on YOLO(Redmon et al. 2016)(Redmon and Farhadi 2018) (Bochkovskiy, Wang, and Liao 2020).

Adding attention mechanisms on top of it can allow the model to focus more on important features. For YoLov5, commonly used attention mechanisms include: SE (Squeeze-and-Excitation) attention module, ECA (Efficient Channel Attention) attention module, CBAM (Convolutional Block Attention Module) attention module, and CA (Channel Attention) attention module. These four types of attention modules only require a small amount of additional computational resources while enhancing the model's feature learning capabilities and optimizing its performance. They can be readily applied as needed.

## Related Work

In the past, surface defect detection on steel relied heavily on manual visual inspection, resulting in low detection rates and an inability to meet production demands. With the gradual maturation of computer vision and image processing technologies, computer vision-based surface defect detection techniques have gradually replaced manual inspection in industrial production. In recent years, breakthroughs

in deep learning and artificial intelligence technologies have led to the emergence of various object detection algorithms, providing more possibilities for industrial surface defect detection.

Therefore, researching the integration of deep learning algorithms with steel surface defects to meet real-time and accuracy production needs is of significant importance for accelerating the advancement of the steel industry and improving our country's industrial capabilities.

Currently, there are two main types of deep learning-based object detection algorithms: two-stage and one-stage object detection algorithms .

Two-stage object detection algorithms, as the name suggests, break down the object detection task into two subtasks. The first subtask involves generating candidate boxes that ideally contain the target objects. The second subtask entails using convolution to extract relevant features from the candidate box regions, which are then fed into a classification network to predict the object's category. These two steps are combined to complete the object detection. Representative networks in this category include Faster R-CNN (Ren et al. 2015), R-FCN (Dai et al. 2016), and Mask R-CNN (He et al. 2017). While two-stage object detection algorithms often offer high precision, their inference process involves complex computations across two stages, making them unsuitable for real-time detection.

One-stage object detection algorithms treat the two subtasks of two-stage algorithms as a single task and employ an end-to-end structure. When an image is input, they directly produce the categories and positions of objects within the image. Representative networks in this category include YOLO and SSD (Liu et al. 2016).

In the subsequent development of models,introduced a lightweight version of Faster R-CNN (Ren, Geng, and Li 2018) They replaced the convolution layers used for feature extraction with depthwise separable convolutions, resulting in a three to fourfold increase in network speed . Additionally, they added center loss to the original loss function to enhance the network's ability to differentiate between different types of defects.

Wang (Wang et al. 2021) input images into an improved ResNet50 model, which included deformable convolution networks (DCN) and enhanced cropping for classifying samples with and without defects. If the probability of having a defect is less than 0.3, the algorithm directly outputs defect-free samples. Otherwise, the samples are further input into an improved Faster R-CNN, which includes spatial pyramid pooling (SPP), enhanced feature pyramid network (FPN), and matrix non-maximum suppression (NMS). The final output is the location and classification of defects or defect-free regions within the samples.

Kou et al. developed an end-to-end defect detection model based on YOLO-V3(Kou et al. 2021) They used an anchor-free feature selection mechanism to choose the ideal feature scale for model training, replacing anchor-based structures to reduce computation time. The model introduced specially designed dense convolution blocks to extract rich feature information, effectively improving feature reuse, feature propagation, and enhancing the network's representational capacity.

While the above-mentioned methods have achieved good results in object detection, there is still room for improvement, especially in the detection of small targets like industrial defects under weak supervision conditions, which are essential to meet practical needs(Zhang, Ding, and Yan 2011).

## Proposed Solution

In our research, we chose YOLOv5 as the base model for industrial surface defect detection. YOLOv5, known for its speed and accuracy in complex scenes, was customized to meet industrial defect detection needs. We introduced a fused attention mechanism to enhance focus on critical regions, aiming to boost defect detection performance. YOLOv5 was selected due to its excellent object detection performance and adaptability. Customized modifications were applied for alignment with our specific task requirements.

### The Network Architecture

YOLOv5, a state-of-the-art object detection algorithm, comprises four key components: the input module, backbone network, neck network, and detection head. Notable modules within YOLOv5 include the Focus module, CBL convolution module, CSP module, and Concatenation module. The Focus module is designed to enhance small object detection, the CBL convolution module integrates Convolutional, BatchNorm, and Leaky ReLU operations for feature extraction, the CSP module aids in effective feature fusion, and the Concatenation module facilitates seamless concatenation of feature maps. YOLOv5 excels in real-time object detection tasks, demonstrating impressive accuracy and efficiency across diverse applications.

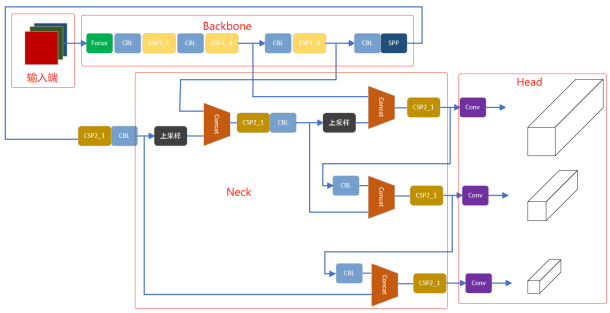


Figure 1: In the network architecture of YOLOv5, the attention module can be inserted into the last layer of the backbone network or into the Neck network to enhance feature fusion capability. In this paper, we choose to insert the attention module after the three-scale feature maps output by the Neck network. After applying the attention mechanism, the feature maps are then fed into the detection head

### Adding Attention Mechanism

Considering the diverse and complex nature of industrial surface defects, traditional detection methods may struggle to capture crucial defect regions accurately. Attention

mechanisms enhance the model's focus on vital areas during learning, improving its perception of defect regions. Our goal is to boost the model's robustness in real industrial settings, allowing more precise detection and localization of surface defects. This provides a reliable solution for industrial quality control under weak supervision. We believe incorporating attention mechanisms enhances industrial surface defect detection algorithms.

1. Squeeze-and-Excitation Network (SENet), introduced by Jie Hu et al., emphasizes channel relations in CNNs. Unlike previous research focusing on spatial aspects, SENet recalibrates channel features by explicitly modeling interdependencies. Despite a slight increase in computational cost, SE blocks notably enhance the performance of state-of-the-art neural networks.

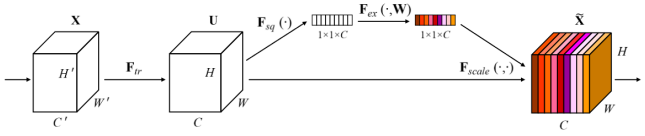


Figure 2: Squeeze and Excitation Network (SENet)

2. Efficient Channel Attention (ECA) Module, presented by Qilong Wang et al. in 2020, enhances channel features through weighted scaling in the feature map. Differing from other attention modules, ECA uses equivalent class convolution to abstract channel relationships, effectively extracting feature importance within each channel. Despite a similar parameter count to the base model, this method yields significant performance improvement.

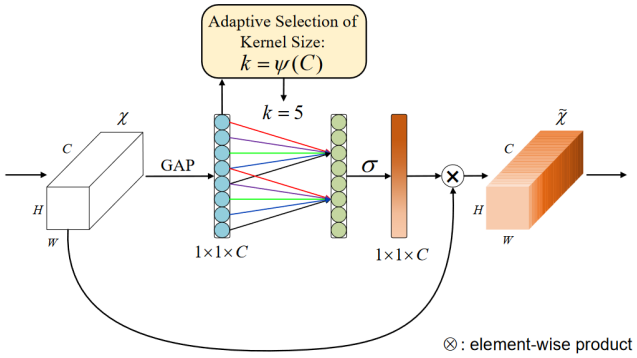


Figure 3: Efficient Channel Attention (ECA) Module

3. Convolutional Block Attention Module (CBAM) combines Channel Attention Module (CAM) and Spatial Attention Module (SAM). Unlike SE and ECA modules, which emphasize channel attention, CBAM uniquely integrates both channel and spatial attention mechanisms. This enhances the model's ability to learn discriminative features for improved object categorization via channel attention and to focus on different image positions through spatial attention. This synergistic approach allows the network better control over feature importance and distinctiveness, facilitating more precise feature learning.

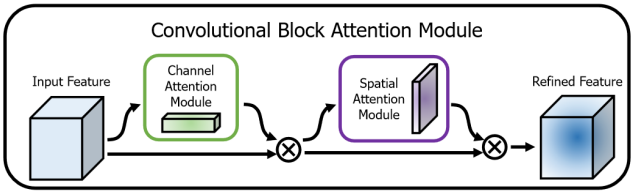


Figure 4: Convolutional Block Attention Module

4. Coordinate Attention (CA) Module, proposed by Qibin Hou et al. in 2021, tackles spatial neglect in channel attention. CA integrates spatial data by breaking down global channel attention pooling into width and height dimensions, embedding position data into channels. This dual-channel approach ensures the model considers both feature importance and spatial relationships. These four modules, while slightly increasing computation, boost feature learning in the model, optimizing performance. They efficiently enhance capabilities with minimal resource consumption, maintaining inference speed. In YOLOv5, insert these modules after the Neck network's three-scale feature maps for improved feature fusion.

## EIoU Loss Function

YOLOv5 employs CIoU (Complete Intersection over Union) as the bounding box loss, incorporating an aspect ratio factor to measure the aspect ratio comprehensively while considering both area and distance. The specific calculation formula is as follows:

$$CIoU_{Loss} = 1 - (IoU - \frac{d_0^2}{d_c^2} - \frac{v^2}{1 - IoU + v}) \quad (1)$$

$$v = \frac{4}{\pi}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w^P}{h^P}) \quad (2)$$

However, $CIoU_{Loss}$ has two issues. First, when the aspect ratio of the predicted box is the same as or proportional to the true box, the penalty term related to v becomes ineffective. Second, according to the gradient formula for calculating aspect ratio, the length and width of the predicted box are inversely proportional, making it impossible to simultaneously increase or decrease both. Therefore, we decided to enhance it using the EIoU (Exponential Intersection over Union) loss, which introduces the length and width information of the target box and is calculated as follows:

$$EIoU_{Loss} = L_{IoU} + L_{dis} + L_{asp} \quad (3)$$

$$EIoU_{Loss} = 1 - IoU + \frac{d_0^2}{d_c^2} + \frac{d_o^2(w^p, w^{gt})}{(w^c)^2} + \frac{d_o^2(h^p, h^{gt})}{h^c)^2} \quad (4)$$

Modifying from CIoU to EIoU aims to boost industrial surface defect detection under weak supervision. EIoU, with an exponential term, heightens IoU calculations, enhancing sensitivity to predicted vs. true box overlap. Due to sparse

annotations in defect detection, models face localization errors. EIoU, as the loss function, targets better accuracy and robustness, especially under weak supervision. This adaptation aims to increase adaptability in intricate industrial scenarios, ensuring a reliable surface defect detection solution.

# Experiments

## Dataset and Parameter Settings

The dataset utilized in this study is the NEU-DET dataset(as shown in Figure 5), curated by Associate Professor Ke-Chen Song from Northeastern University. The dataset comprises six typical surface defects of steel, including rolled-in scale, patches, crazing, pitted surface, inclusion, and scratches. Each class consists of 300 images, resulting in a total of 1800 images. The pixel size of each image is 200x200, and each image is associated with a corresponding label file containing information about the defect's position and size. The

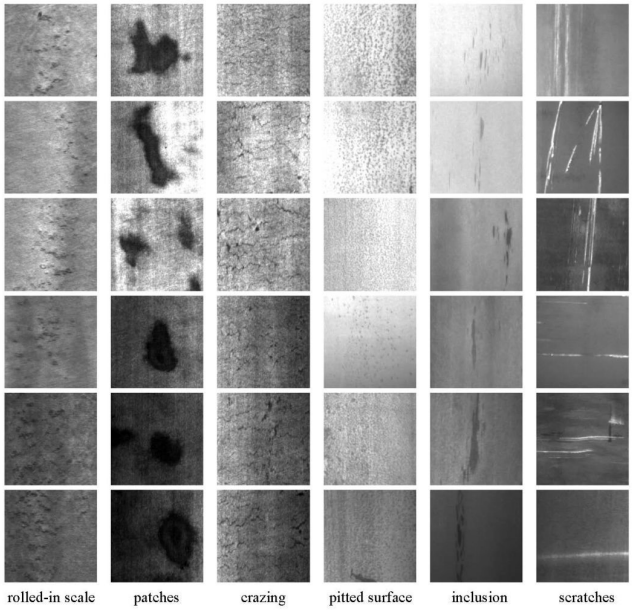


Figure 5: The presentation of some aspects of the NEU dataset reveals that there are significant visual differences among intra-class defects. For example, scratches (last column) can be horizontal, vertical, or diagonal. Similarly, inter-class defects share similarities, such as scale, crack, and pitted surface. Additionally, the grayscale variations of intra-class defect images are influenced by lighting and material changes

experimental setup employs a batch size of 16 and image dimensions of 200x200. During actual training, the images are dynamically resized to 224x224. This resizing is necessary as the YOLOv5 model's feature extraction backbone undergoes five downsampling steps, requiring a 32x reduction. Thus, the input images are automatically resized to multiples of 32 for efficient model computation. The training consists of 100 epochs with an initial learning rate of 0.01, utilizing the Adam optimization method.

## Model Performance Evaluation Metrics

This paper employs precision (P), recall (R), and mean average precision (mAP) as the evaluation metrics for assessing the model's performance. The specific calculation formulas are as follows:

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$AP = \int_0^1 P(r)dr \tag{7}$$

$$mAP = \frac{\sum_{i=0}^{n} AP(i)}{n} \tag{8}$$

P (Precision) denotes the proportion of samples predicted as positive by the model that are truly positive, while R (Recall) represents the proportion of all truly positive samples that the model successfully predicts as positive. mAP (mean Average Precision) is the average value under the Precision-Recall curve, providing a comprehensive assessment of both the model's accuracy and recall.

## Results Analysis

In the YOLOv5s model results after 100 epochs are shown. Loss analysis reveals bounding box loss ($box_{loss}$) and category loss ($cls_{loss}$) converge swiftly, with bounding box loss slowing around 50 epochs and category loss converging around 25 epochs. However, confidence loss ($obj_{loss}$) converges slowly, displaying oscillations. Training set confidence loss decreases after 100 epochs, while validation set confidence loss stabilizes between 50 and 100 epochs, suggesting a risk of overfitting with prolonged training.
The reason for this situation may be due to the inter-class similarity in the dataset, where the single-channel grayscale images have similar features among different categories. This can result in the model being influenced by a lot of noise and not effectively learning the distinctive key features of each class

Figure 6 displays YOLOv5s' P-R curves after 100 epochs, yielding an mAP of 0.766, the mean across six defect types. In-depth analysis reveals subtle features in crazing and rolled-in scale, challenging the model and causing performance decline, highlighting a significant issue in the steel surface defect dataset.
Upon careful analysis, it was noticed that two types of defects, namely silver lines and rolled scales, have less distinct features compared to the original background. The model might not have effectively learned the characteristics of these two defect classes. This poses a challenge in the steel surface defect dataset. In future research, a possible approach could be to focus on these two defect classes and further investigate how to extract their specific shapes and patterns. By enhancing the model's ability to differentiate these two classes, the overall performance of the model can be improved.
In addition to presenting the training results and P-R curves

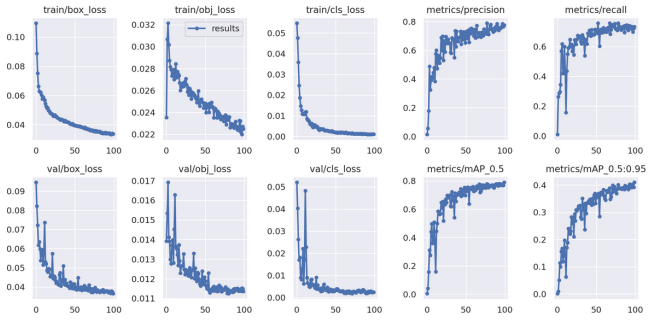


Figure 6: The model's accuracy and recall gradually increase and converge over time. Around the 25th epoch, the growth rate of both accuracy and recall starts to slow down. The mean Average Precision (mAP) exhibits a similar growth curve, eventually stabilizing at around 0.8. Overall, the model performs well on the steel surface defect dataset.

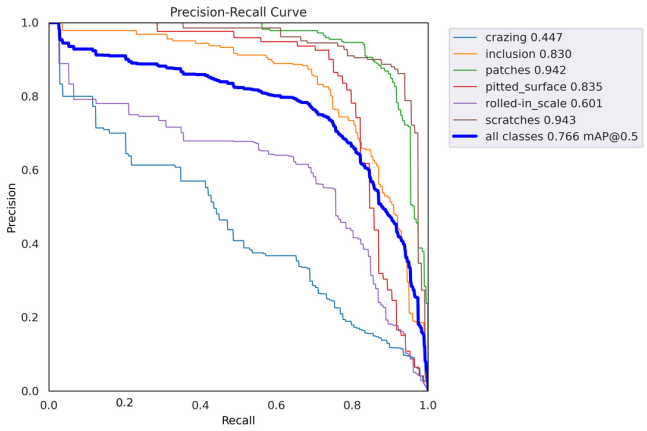


Figure 7: Among spots, pitted surface, inclusions, and scratches, individual mAPs are 0.942, 0.835, 0.830, and 0.943, surpassing the overall average. Conversely, crazing and rolled-in scale perform poorly, with mAPs of 0.447 and 0.601, impacting overall model performance.

of the YOLOv5s base model, this paper introduces enhancements to YOLOv5s. SE, ECA, CBAM, and CA attention mechanisms are added after the multi-scale feature fusion module in the neck, which outputs feature maps at three scales. Comparative experiments are conducted, and the EIoU is employed as the bounding box loss to further explore the model's performance.

In Table 1, attention mechanisms enhance model performance. Despite SE module's notable increase in layers and parameters, it provides a modest 0.2% improvement. ECA, with comparable parameters, achieves a more substantial 1.2% boost. CBAM, with the largest parameter increase, leads to a 1.7% improvement. CA, with a modest parameter rise, exhibits the most significant improvement at 2.4%. Notably, all attention modules maintain operational speed around 16.0 GFLOPs, ensuring enhanced performance without compromising computational efficiency.

Table 1: Model Performance Comparison

| Model | Network Layers | Parameters | GFLOPs | mAP@0.5 |
|---|---|---|---|---|
| base | 270 2 | 7035811 | 16.0 | 0.766 |
| +SE | 2917 | 7078819 | 16.0 | 0.768 |
| +ECA | 282 | 7035820 | 16.0 | 0.778 |
| +CBAM | 303 | 7079113 | 16.1 | 0.783 |
| +CA | 300 | 7071491 | 16.0 | 0.790 |

## Conclusion

In the context of surface defects in steel, there are significant similarities between defect categories and substantial variations within the categories. Additionally, the imaging of defects is often influenced by different materials and lighting conditions, resulting in varying image quality. Furthermore, in real industrial production lines, steel production occurs at a certain speed, necessitating defect detection algorithms to deliver both precision and real-time performance to meet the practical needs of industrial production. The proposed improvements to the single-stage object detection algorithm YOLOv5 not only enhance precision but also consider inference speed, offering significant advantages for deployment in real-world industrial environments.

1. In actual industrial production scenarios, controlling the rate of substandard products results in a severe imbalance between positive and negative data samples, with a scarcity of defect samples and a large number of unlabeled images. Under weakly supervised conditions, it becomes more effective to utilize data samples.
2. The addition of attention mechanisms can enhance the model's perceptual capabilities and improve its ability to learn features, enabling the deep learning of unique characteristics of each defect and distinguishing between them. This reduces the impact of image noise and environmental factors during actual detection.
3. The use of EIoU is more in line with the practical situation of defect detection, enhancing the model's robustness and enabling it to perform pre-detection tasks effectively and generate reasonable prediction boxes when confronted with different datasets or deployed on different production lines.

The experimental results on the NEU-DET steel surface defect dataset from Northeastern University demonstrate that various improvement modules have enhanced the model's performance, with the highest improvement reaching 2.4%, resulting in the best mAP value of 79.0%. Meanwhile, the increase in model parameters is minimal, and the computational workload of the model remains largely unchanged. This validates that by only adding a portion of computational resources, the model's performance can be improved without affecting the inference speed.

## References


Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Kou, X.; Liu, S.; Cheng, K.; and Qian, Y. 2021. Development of a YOLO-V3-based model for detecting defects on steel strip surface. *Measurement*, 182: 109454.

Lee, S. Y.; Tama, B. A.; Moon, S. J.; and Lee, S. 2019. Steel surface defect diagnostics using deep convolutional neural network and class activation map. *Applied Sciences*, 9(24): 5449.

Li, J.; Su, Z.; Geng, J.; and Yin, Y. 2018. Real-time detection of steel strip surface defects based on improved yolo detection network. *IFAC-PapersOnLine*, 51(21): 76–81.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37. Springer.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, Q.; Geng, J.; and Li, J. 2018. Slighter Faster R-CNN for real-time detection of steel strip surface defects. In *2018 Chinese Automation Congress (CAC)*, 2173–2178. IEEE.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Wang, S.; Xia, X.; Ye, L.; and Yang, B. 2021. Automatic detection and classification of steel surface defect using deep convolutional neural networks. *Metals*, 11(3): 388.

Wei, R.; Song, Y.; and Zhang, Y. 2020. Enhanced faster region convolutional neural networks for steel surface defect detection. *ISIJ international*, 60(3): 539–545.

Zhang, X.; Ding, Y.; and Yan, P. 2011. Vision inspection of metal surface defects based on infrared imaging. *Acta Optica Sinica*, 31(3): 0312004.