Review of Steel Surface Defect Detection and Lightweighting Methods

Miao Chen
Jilin International University - School of Artificial Intelligence, Changchun, China
**Corresponding Author*

Abstract: Surface defects of steel materials affect product quality and industrial safety. High-precision and high-efficiency detection of surface defects in steel is a necessary condition for the high-quality development of the steel industry. Therefore, this paper reviews the research work related to steel surface defect detection and lightweighting. Starting from the data sets, it then elaborates on the defect development of traditional detection technologies, followed by a summary of detection based on machine learning and deep learning. The optimization approaches and lightweighting technologies represented by the YOLO series, such as structural optimization, model compression, and auxiliary strategies, are emphasized and elaborated on as the key points. Moreover, YOLO series model cases are used for illustration, and the effects after different improvement methods in the three cases are analyzed and compared. The YOLOv8-CSG has a better balance: the computational cost is reduced by 37% and the parameter quantity is reduced by 35.2%; for the improvement of the YOLOv8 model, the optimal compression is mentioned: the model size is reduced by 44%, and the lightweighting methods have improved the detection accuracy, detection efficiency, and real-time performance. Finally, the relevant technologies for detecting surface defects of steel were briefly summarized, and the future development of these technologies was also prospected.

Keywords: Steel Surface; Defect Detection; Lightweighting; Deep Learning

1. Introduction

China is the world's largest steel producer. In 2024, the total global crude steel output was 1.885 billion tons, and China's crude steel output accounted for 53.3%, occupying the vast majority of the global steel production. Therefore, the demand for steel in China is large and extensive. However, surface defects of steel (such as cracks, scratches, oxide scales, inclusions) are still one of the important factors affecting the qualification rate of products. Defects such as cracks, inclusions, bubbles, and indentations caused during the production processes of smelting, rolling, and heat treatment not only reduce the mechanical properties and corrosion resistance of steel, but also may cause other risks. According to statistics, the scrap rate of steel due to surface defects is approximately 5%—8%, which causes huge economic losses and social risks. Therefore, efficient and precise surface defect detection is an inevitable requirement for the high-quality development of the steel industry.

During the development of detection technology, traditional detection methods mainly relied on manual visual inspection, which had problems of low efficiency and high detection errors. Later, methods such as eddy current detection and eddy current testing were still unable to meet the requirements of automation and intelligence in modern steel production. The detection technology based on deep learning did indeed significantly improve the situation, but it also faced issues such as large model parameters, high computational costs, and difficulty in deployment. Therefore, while ensuring the detection accuracy, optimizing the model through lightweighting technology to reduce computational costs and storage requirements has become a core challenge in promoting the practical application of steel surface detection technology.

This article systematically reviews the development of steel surface defect detection technology, as well as the optimization methods and application scenarios of lightweight technology improvements. The aim is to provide a reference basis for the intelligent and efficient deployment of defect detection systems in the steel industry.

2. Overview of Steel Surface Defect Detection Technology

2.1 Dataset

2.1.1 NEU dataset

The NEU dataset [1] is a surface defect database publicly released by Northeastern University. This dataset consists of a total of 1800 grayscale images, covering six types of steel surface defects: cracks (Cr), inclusions (In), patches (Pa), pitted surfaces (PS), rolled oxide scales (RS), and scratches (Sc). As shown in Figure 1.

2.1.2 Severstal steel defect detection dataset

The Severstal Steel Defect Detection dataset is provided by Severstal Company and is used for detecting and classifying surface defects of steel strips. This dataset contains 6666 high-resolution images of the surface of steel strips, and it includes four types of defects. As shown in Figure 2.

2.1.3 GC10-DET dataset

The GC10-DET dataset [7] is a surface defect dataset collected from real industrial environments, containing ten types of surface defects such as punching (Pu), weld seam (WI), crescent-shaped gap (Cg), water spot (Whater Sport), streak (Ss), oil spot (Os), inclusions (In), rolling pit (Rp), crease (Cr), and waist crease (Wf), with a total of 3,570 grayscale images. Some labeled images are shown in Figure 3.

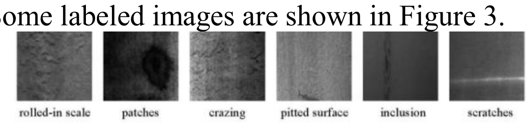

Figure 1. NEU Dataset

Figure 2. Severstal Steel Defect Detection Dataset

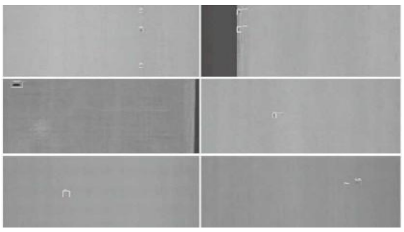

Figure 3. Partial Labeled Image of GC10-DET

2.2 Traditional Detection Technology

2.2.1 Traditional manual visual inspection

Manual visual inspection is the earliest and most widely used method for detecting defects in steel. Through manual visual observation or by using cameras and image processing technology, defects on the steel surface (such as scratches, pits, corrosion, etc.) are identified, including processes such as image preprocessing, feature extraction, defect detection and classification. [2] Traditional manual inspection is very slow, takes a long time, and requires a lot of manpower. During the inspection process, due to various reasons, the efficiency and quality of the inspection may decrease, the inspection quality may vary, and it may not meet the standards, thereby affecting the inspection quality. Moreover, in harsh inspection environments, it may cause misjudgment by the inspectors, affecting the correctness of the inspection. In addition, long-term exposure to such environments may cause harm to the physical health of the inspectors.

2.2.2 Magnetic flux leakage detection method

As a technical means of non-destructive testing, [3] this technology is applied to the defect detection in the inspection of metal materials. The application of magnetic flux leakage detection in steel materials of the steel industry can detect cracks, pores, etc. existing on the surface and inside of the steel. Magnetic flux leakage detection is based on the change of magnetic induction intensity to detect defects inside or on the surface of the material. As shown in Figure 4. Using the method of magnetic flux leakage detection to detect surface defects of steel is an effective and non-destructive means. Analyzing the changes in its magnetic field helps to ensure the safety and reliability of the steel structure.

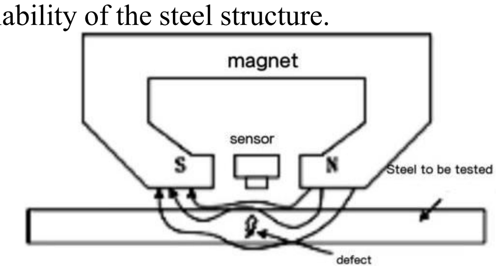

Figure 4. Schematic Diagram of Leakage Flux Detection Principle

2.2.3 Eddy current testing method

The eddy current detection technology is based on the principle of electromagnetic induction [4]. As shown in Figure 5. It detects whether there are defects on the surface by sensing the changes of eddy currents induced in the conductive material. First, an alternating voltage is provided to a probe, which is a coil. A changing magnetic field is generated around the coil, and this changing magnetic field can produce a changing current inside the steel piece. This changing current is called eddy current. The magnitude and distribution of the eddy current are related to the conductivity, thickness, and temperature of the material (such as steel), as well as the defects on the surface (such as cracks or pores, etc.). Therefore, after the magnetic field is emitted by the probe and eddy currents are generated on the steel body, it is possible to know the state of the steel body and whether there are any problems on the surface.

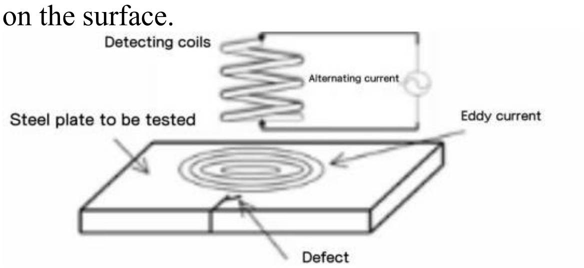

Figure 5. Schematic Diagram of Eddy Current Detection Principle

2.3 Machine Learning and Deep Learning Detection Methods

In recent years, computer vision and deep learning technologies have gradually been applied to the automatic detection of steel surface defects. With the introduction of the AlexNet network, convolutional neural networks (CNNs) have also been widely used in the field of computer vision, featuring high efficiency, high accuracy, and strong robustness. After the R-CNN object detection algorithm was introduced in 2014, the Faster-RCNN two-stage object detection algorithm emerged in 2015, achieving an end-to-end object detection framework and becoming a classic benchmark algorithm in the field of object detection. Although the object detection algorithms have been continuously improved and advanced, there are still some areas that can be optimized for further work, and some scholars have successively made improvements to this algorithm. For example: Yu Qinghua [5] improved Faster-RCNN by using the K-means algorithm to automatically generate 5 sets of prior boxes to enhance the regression ability for multi-scale defects. At the same time, it adopted stepwise training to speed up the process, used histogram equalization to enhance the test image, and combined Softmax with Center Loss as the joint loss function. The experiments showed that the improved algorithm could increase the average accuracy rate of steel surface defects to 75.1%, and also increase the detection speed by 1 frame/s (from the original 18 frames/s to 19 frames/s). However, this algorithm only improved its detection speed, and did not have a significant breakthrough in detection accuracy.

Traditional machine learning methods can perform steel surface defect classification and detection by manually extracting features. These algorithms are used to identify defects such as scratches, pits, and rust.

3. An Overview of Lightweight Methods for Detecting Defects on Steel Surfaces

In the process of intelligent steel manufacturing, although the steel surface defect detection technology based on deep learning has solved many problems of low accuracy, the parameters of the mainstream models are quite large and they rely on high-performance GPUs, making it difficult to deploy on devices. In industrial sites, real-time detection is required and the cost of controllable hardware is a concern. Therefore, lightweight methods have become a key breakthrough path.

3.1 Core Objectives and Technical Framework for Lightweighting

3.1.1 Lightweighting objective

The lightweight requirements for detecting surface defects of steel have distinct industrial characteristics. Firstly, it is necessary to maintain accuracy, reduce the rate of missed detections and false detections, and solve problems such as adapting to tiny cracks; secondly, it is necessary to improve efficiency, reduce the number of model parameters and computational load; finally, it is necessary to adapt to various complex environments and types of steel.

3.1.2 Lightweight technical approach

(1) Bottom layer (model structure reconfiguration): Use lightweight architectures such as MobileNet, ShuffleNet, or specialized structures.

The MobileNet series reduces the computational load through depthwise separable convolutions. For instance, Hu Mingqi et al. [6] addressed the issue that the accuracy and lightweighting of YOLOv8n could not be achieved simultaneously in the defect detection on steel surfaces, and they constructed the YOLOv8n-MDC model. Firstly, WIoU was used to replace the built-in IoU of YOLOv8n. A non-monotonic focusing mechanism was added, which enhanced the model's ability to accurately locate defect positions and improved the robustness of the entire model. In complex backgrounds, the model reduced the occurrence of false detections and missed detections. Secondly, the MobileNetV3 network is used to replace the original model's backbone feature extraction block to achieve model lightweighting. Then, the original network modules are replaced with DW convolution + C3Ghost module. According to the verification on the NEU-DET steel surface defect dataset, the performance of the YOLOv8n-MDC model has achieved significant results: Firstly, the average precision mean reaches 81.3%, which is 5 percentage points higher than that of the original model. Secondly, the parameter quantity is only 1.02M, which is 33.9% of the original model; the computational power is 2.1 GFLOPs, which is 25.9% of the original model, and it also meets the requirements of industrial scenarios.

For another example, Jiang Bo et al. [8] improved the original YOLOv5s model by replacing the main backbone extraction network of the original model with the MobileNetv3-Small network. This approach reduces the complexity of the model at its source and introduces the weighted bidirectional feature pyramid network (BiFPN) to enhance the feature capture of defects of various sizes, thereby improving the detection accuracy and robustness. Adding the convolutional block attention module can reduce the missed detection of small targets. Finally, the K-means++ algorithm is used to cluster the prior boxes to optimize the efficiency of candidate box matching. Based on the NEU-DET steel defect dataset and NVIDIA 1080Ti hardware verification, the accuracy has been improved, with mAP@0.5 reaching 77.2%, which is 3.90% higher than the original YOLOv5s. The detection ability for small-scale and multi-type defects has been significantly enhanced; the parameter quantity has been reduced by 58.6%, the model size has been decreased by 34%; the detection speed has reached 102FPS, which is 29.7% higher than the original model, and meets the real-time detection requirements.

(2) Middle-level (Model Compression): Conduct relevant work on the high-precision model after training, such as channel pruning, quantization, and knowledge distillation. For example, Ma Yanting [9] in the proposed MT-YOLOv5 model, based on the accurate detection accuracy, adopted a "compression + acceleration" dual strategy to optimize model deployment: Firstly, through pruning, significantly reduce the number of parameters and computational cost; laying the foundation for the adaptation of edge devices; then, using Tensor RT for acceleration, combined with low-precision data types and device deployment, while considering accuracy, real-time performance, and low resource consumption, provided a high-level method for steel surface defect detection.

(3) Upper layer (algorithm strategy adaptation): Due to the weak feature extraction capabilities of the models at the bottom and middle layers, the methods for compensating for the accuracy loss caused by model lightweighting using upper-layer technologies include: attention mechanism fusion, multi-scale feature fusion, data augmentation, and transfer learning, etc.

4. Application Cases and Effect Evaluation

4.1 Case Analysis of Typical Improvement Model

An analysis was conducted on three improved algorithms for detecting surface defects of the YOLO series of steel materials.

4.1.1 STC-YOLOv8 model

The STC-YOLOv8 model was proposed by Huang Aoguo et al. [10] to address the issues of high false negatives and false positives and low accuracy and efficiency in traditional detection due to the "small size and complex background" of steel surface defects. This model is an improved version of YOLOv8. Firstly, the original convolution layers were replaced with spatial attention convolution (SAConv) to enhance the model's ability to extract features for small target defects. Secondly, the content-aware upsampling (CARAFE) was used to replace the original upsampling method, improving the model's accuracy in extracting defect edge features and similar features, and adapting to defect recognition in complex backgrounds. A multi-branch structure with re-parameterized modules was introduced, and a transfer learning strategy was adopted to optimize the accuracy of classification and localization in the detection process, enhancing the model's generalization and stability.

4.1.2 YOLOv8n-CSG model

The YOLOv8n-CSG model is a lightweight algorithm for improving the YOLOv8n proposed by Zhao Baoting et al. [11]. Firstly, the context-guided module (CG block) is introduced to enhance the feature capture ability, and the C2f_CG module is designed; secondly, the star-shaped network module is added to optimize the processing of minor defects, and the C2f_Star model is constructed; finally, a lightweight detection head is designed. The GSE_Detect detection head integrating GSConv (ghost shadow mixing convolution) and the efficient multi-scale attention mechanism is designed, which retains the original detection efficiency while reducing the model complexity and achieving lightweighting.

4.1.3 Improvement of the YOLOv8 algorithm

Chang Le et al. [12] addressed the core issues of insufficient accuracy and limited speed in the detection of surface defects of steel materials. They improved the YOLOv8 algorithm by optimizing it through three modules: information retention, lightweight design, and accuracy enhancement.

Firstly, the C2f-AKConv module is constructed by combining AKConvBottleneck and C2f modules. This not only reduces the information loss during the feature extraction process but also decreases the computational load of the model. Secondly, a new downsampling module, SCDown, is designed. By optimizing the information transmission path, it effectively enhances the retention ability of key features during downsampling, strengthening the foundation for detecting subtle defects. In the feature fusion stage, a scale feature fusion module (CCFM) is introduced to enhance the feature correlation of multi-scale defects, thereby improving the detection accuracy.

4.2 Effect Appraisal

4.2.1 Precision performance

Among the three steel surface defect detection algorithms, the STC-YOLOv8 has good adaptability across different datasets: the mAP@0.5 in the NEU-DET dataset reaches 82.4%, and in the GC10-DET dataset it reaches 66%; the YOLOv8n-CSG model has a significant advantage in accuracy: the mAP0.5 reaches 76.8%; in the improved YOLOv8 model, the accuracy has steadily improved: the F1 score has increased by 3.6%, and the mAP@0.5 has increased by 3.3%.

4.2.2 Lightweighting capability

In the lightweight capability analysis, YOLOv8-CSG achieves better balance: the computational cost is reduced by 37% and the parameter quantity is decreased by 35.2%; in the improvement of the YOLOv8 model, it is mentioned that the compression is optimal: the model size is reduced by 44% and the parameters are significantly decreased.

5. Conclusion and Prospect

5.1 Research Summary

This study reviews the methods for detecting surface defects of steel and lightweighting techniques. It progresses from traditional manual detection or machine learning detection to deep learning detection such as the YOLO series. The lightweighting methods address the above issues from three aspects: structural optimization, model compression, and auxiliary strategies. Among the existing improved algorithms, they have achieved varying degrees of accuracy improvement and reduction in model parameters, and can partially adapt to actual industrial production scenarios.

5.2 Future Expectations

The current lightweight model for detecting surface defects of steel materials has the problems of insufficient understanding of technical details in extreme environmental conditions and poor applicability in small sample defect scenarios. If we can be flexible and adaptable based on actual circumstances, and use devices such as cameras and infrared sensors to collect data, and then have this data be processed and analyzed by various servers and equipment to detect surface defects of steel strips, this will better meet the detection requirements for various materials and defects.

References

[1] Song Yubin, Kong Weibin, Chen Xi, et al. Review on Research on Detection of Surface Defects of Steel [J]. Software Review, 2024, 23(03): 203-211.

[2] Li Yanshun. Research on Steel Surface Defect Detection Based on Deep Learning [D]. Shandong Jiaotong University, 2024.

[3] Zhao Yufeng. Steel Surface Defect Detection Based on Deep Learning [D]. Shenyang University of Technology, 2024.

[4] Li Yuandong, He Zhiming. An Improved YOLOv8 Algorithm for Steel Surface Defect Detection [J]. Software Review, 2024, 23(11): 200-205.

[5] You Qinghua. Research on Steel Surface Defect Detection Method Based on Deep Learning [D]. Wuhan Textile University, 2021.

[6] Hu Mingqi, Chen Huiming, Xu Wei, et al. Lightweight YOLOv8 Steel Surface Defect Detection Incorporating MobileNetv3 [J]. Science & Technology & Engineering, 2025, 25(16): 6831-6840.
[7] Ma Rui, Fu Xingkui, Zheng Guijun, et al. Steel Surface Defect Detection Based on Improved YOLOv5 Algorithm [J]. Internet of Things Technology, 2025, 15(03): 39-43.
[8] Jiang Bo, Wan Yi, Xie Xianzhong. A Lightweight Steel Surface Defect Detection Model Based on Improved YOLOv5s [J]. Computer Science, 2023, 50(S2): 271-277.
[9] Ma Yanting. Research on Steel Surface Defect Detection Technology Based on Deep Learning [D]. Hebei University of Technology, 2023.
[10] Huang Aoguo, Luo Xiaoling, Pan Xin. STC-YOLOv8: A Detection Method for Steel Surface Defects in Complex Backgrounds [J]. Journal of Inner Mongolia Agricultural University (Natural Science Edition), 2024, 45(05): 67-76.
[11] Zhao Baiting, Zhang Min, Jia Xiaofen. YOLOv8n-CSG: A Lightweight Algorithm for Detecting Surface Defects of Steel Materials [J/OL]. Journal of Electronic Measurement and Instrumentation, 1-13 [2025-10-11]
[12] Chang Le, Liu Guiru, Wang Lu Lin, et al. A Lightweight YOLOv8 Algorithm for Detecting Surface Defects of Steel Materials [J/OL]. Journal of Tianjin University of Technology, 1-10 [2025-10-11]