Available Online: 25 June 2025 DOI: 10.54254/2977-3903/2025.24461

# A review of deep learning-based detection and segmentation of steel surface defects: advances in resolving intra-class differences and inter-class similarities

*Zibo Zhao*

Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

15639159661@163.com

**Abstract.** Steel surface defect detection is a key part of the production process in the steel industry. The traditional manual inspection methods are inefficient and costly. With the rapid development of deep learning technology, automatic detection and segmentation of steel surface defects based on deep neural networks has received widespread attention and demonstrated good performance in several real-world scenarios. However, challenges remain due to the obvious inter-class similarity and intra-class variation problems in steel defect images, impacting accuracy and robustness. This paper systematically summarizes the representative researches on solving the inter-class similarity and intra-class variation problems in recent years, and focuses on analyzing the innovations of different methods in network structure design. In addition, this paper also discusses the shortcomings of the current research, and proposes a new idea of fusion of mainstream modeling method. By employing a subcomparison module to enhance feature similarity within classes and differentiate across classes, followed by pyramid feature fusion to optimize computational efficiency, this study aims to advance high-precision intelligent recognition of steel surface defects. This study reveals that the approach not only addresses existing challenges but also provides a foundation for future advancements in steel defect detection technologies.

**Keywords:** convolutional neural networks, intra-class differences, surface defect detection segmentation, inter-class similarity

## 1. Introduction

Surface defect detection is an important part of guaranteeing the quality of industrial products, especially in the steel manufacturing process is of key significance. High-quality steel is widely used in construction, transportation, machinery manufacturing and other fields, and its quality is directly related to the safety and stability of the engineering structure. Experiments have shown that surface cracks with a depth of 0.1mm can reduce the fatigue strength of Q345 low alloy steel by 22%-35% [1]. If there are defects on the surface of steel, it may not only weaken its mechanical properties and service life, but also cause engineering accidents in serious cases, resulting in casualties.

Traditional steel defect detection mainly relies on manual visual inspection or photoelectric detection. Manual inspection is inefficient and subjective, which is difficult to meet the needs of large-scale industrial production; while photoelectric detection requires high equipment accuracy and is greatly affected by the environment, resulting in insufficient stability of the detection results [2]. In recent years, with the development of deep learning technology, the automated defect detection method based on Convolutional Neural Network (CNN) has become a research hotspot, which provides a new path to improve the detection accuracy and efficiency.

Steel surface defect detection faces two core challenges: intra-class variability and inter-class similarity. Intra-class variability is manifested by significant variations in the size, shape, texture, and illumination conditions of defects (e.g., scratches) in the same class, leading to reduced model generalization capability and blurred decision boundaries. This variation makes it difficult for single-scale feature extraction to capture features of both small defects (easily overwhelmed by downsampling) and large defects (insufficient sensory field). Inter-class similarity, on the other hand, is reflected in the high degree of overlap in texture, grayscale, and edge morphology of different classes of defects, resulting in convergence of high-level feature representations and insufficient discriminative information. To cope with these challenges, model design needs to focus on multi-scale feature extraction, contextual modeling, and enhanced inter-class difference perception to improve the accuracy and robustness of detection [3, 4].

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Based on a systematic review of relevant literature, this paper introduces the current mainstream steel defect detection methods. It also analyzes their structural design and performance in solving the intraclass variability and interclass similarity, and summarizes the effective strategies. This paper aims to provide theoretical support and practical guidance for the construction of a more accurate and efficient steel defect detection model in the future.

## 2. Commonly used data sets for steel surface defect detection

High-quality labeled datasets are crucial for promoting steel surface defect detection research. Currently, the widely used datasets in the research field mainly include the NEU dataset released by Northeastern University and the Severstal Steel Defect dataset provided by Severstal Company in Russia [5, 6]. Both of them have significant differences in terms of image source, data size, defect type and task applicability, representing typical data characteristics in experimental environments and industrial scenarios, respectively.

The NEU surface defects dataset, constructed by Kechen Song's team, covers six typical steel defect types. The dataset contains a total of 1,800 images, 300 for each type of defects, with 200×200 pixel grayscale images with uniform resolution and background. The advantages of this dataset include accurate defect labels, balanced sample categories, and standardized format, which is suitable for baseline evaluation and rapid prototype validation of the algorithm performance.

The Severstal Steel Defect dataset is derived from actual steel production lines and is published by Severstal in Russia, reflecting the needs of defect detection in a real manufacturing environment. The images in this dataset are RGB color images with a resolution of 1600×256 pixels, covering production images under a wide range of equipment and process conditions. The training and test sets are approximately 12,568 and 5,578 images in size, respectively, and support defect mask labeling at the pixel level for tasks such as image segmentation and target detection. The advantages of this dataset include high image quality, large sample size, and rich defect types, which are highly representative of the reality and help to evaluate the robustness and practicality of the depth model under the conditions of complex backgrounds and irregular defects.

## 3. Response to the problem of intra-class variability and inter-class variability

### 3.1. Model design and optimization strategies for intraclass variability

The significant variation in target size requires the network to take into account both shallow and deep features in the feature extraction process. However, traditional single-scale feature extraction networks are difficult to effectively handle features at all scales. For this reason, Lin et al. proposed the Feature Pyramid Network (FPN) [7]. FPN realizes the effective fusion of features at different semantic levels by constructing bottom-up feature extraction paths and top-down feature enhancement paths, and introducing lateral connections.

On the basis of FPN, ETDNET further proposed Channel-Modulated Feature Pyramid (CM-FPN) to dynamically regulate the multi-scale feature fusion process by extracting channel-level contextual features through a Global Context Block (GC-Block) and generating an attention weight map [8]. It achieved 46.7% mAP and 80.2% AP50 on the NEU-DET dataset, demonstrating excellent detection performance. Nevertheless, more validation is still required to assess the method's efficacy on the Severstal dataset, which has a high number of data and intricate fault patterns. This is mostly hindered by the inconsistent labeling of this dataset. The experimental results show that ETDNet is particularly suitable for scenarios with high-precision detection and significant defect scale variations, but there is still room for improvement when dealing with complex backgrounds or category-unbalanced data.

PGA-Net A Pyramid Feature Fusion (PFF) module is proposed for the problem of defect detection with high intra-class variability [4]. The module extracts multi-scale features from five stages of the VGG-16 backbone network and realizes feature fusion by downsampling, upsampling and 1×1 convolution, which effectively solves the problem of spatial inconsistency caused by scale variability. Experiments show that PGA-Net achieves 82.15%, 74.78%, 71.31%, and 79.54% mIoU on NEU-Seg, DAGM 2007, MT_defect, and Road_defect datasets, respectively, demonstrating excellent generalization ability. However, the method relies on the VGG-16 network resulting in high computational complexity, and its performance in real-time applications and category imbalance scenarios still needs further validation.

Zhang et al. proposed an improved network based on Mask R-CNN, in which the Self-Contrastive Learning (SCL) module generates two positive sample views of the same image through data augmentation. This approach maximizes the similarity between different views of the same image while minimizing the similarity between different images [9]. It allows the model to better capture different manifestations of defects in the same class, thus enhancing robustness to intra-class variability. Compared with the original Mask R-CNN, the improved network improves in both the average bounding box accuracy (APbbox) and average segmentation accuracy (APsegm) metrics, reaching 66.75% and 63.43%, respectively, which demonstrates its higher accuracy and robustness in the task of detecting defects on steel surfaces. The method performs well in dealing with the intra-

class variability problem; however, it requires high data quality and diversity, making it suitable for scenarios with ample data and a demand for high-precision detection.

MA-YOLO proposed an improved multi-scale Feature Fusion Network (FFN) for the intra-class variability problem [10]. This network enhances the transfer of deep feature information through cross-scale feature connectivity and weighted feature fusion, and dynamically adjusts the contribution of feature maps at different scales, thus improving the detection of multiscale defects. Compared with the baseline YOLOv5s, MA-YOLO's mean average precision mean (mAP50) on AL_Dataset is significantly improved by 2.9% to 88.1%, while the model increases only 0.4M size with a frame rate (FPS) of 34, which meets the real-time requirements of industrial inspection. MA-YOLO effectively enhances the ability to deal with class intra-variance problem, which is especially suitable for industrial scenarios that require high accuracy and real-time detection. However, the lightweight design of MA-YOLO gives it an advantage in resource-constrained environments, but it may face computational resource bottlenecks when dealing with large-scale datasets.

## 3.2. Model design and optimization strategies for intraclass variability

ETDNet proposes a Task-Oriented Decoupled Detection Head (TOD Head), which treats the classification and regression tasks separately in order to solve the problem of inconsistent representations between classification and regression tasks [8]. In the classification branch, ETDNet introduces the Global Feature Refinement Module (GFR). The GFR module models the information at each location in the feature graph across channels and locations through global average pooling and multi-scale context enhancement mechanisms to obtain more semantically cohesive category features. This design can highlight the areas of category distinctiveness and effectively reduce the interference of intra-class morphological differences on classification accuracy.

FDSNet proposes a solution based on the global context attention mechanism, which enhances the model's ability to model long-distance dependencies through the Global Context (GC) module to improve the performance of distinguishing inter-class similarity defects [11]. The module extracts global feature vectors using global average/maximum pooling and dynamically adjusts the feature representations through attention weights, enabling the model to focus on key contextual information and effectively mitigate the interference caused by complex context and inter-class similarity. Experimental results show that FDSNet achieves 78.8%, 63.9%, and 90.2% mIoU on NEU-Seg, Magnetic-tile-defect, and MSD datasets, respectively, while maintaining a real-time processing speed of 135.0-186.1 FPS, which demonstrates an excellent balance of accuracy and efficiency. However, its performance advantage is highly dependent on the effectiveness of global context modeling, which is relatively limited in scenarios with simple backgrounds or significant category differences.

Zhang et al. proposed the Angular Margin Penalty (AMP) module in an improved network based on Mask R-CNN. This module enhances the differentiation of features between classes by introducing angular edge m in the feature space [9]. The AMP module is based on the ArcFace loss function, which makes the features of the same class more compact and the features of different classes more separated by adjusting the angle between the feature vectors and the weight vectors. This design effectively improves the model's ability to learn defective features from different categories.

PGA-Net extracts global context information from low-resolution fused feature maps through global pooling and upsampling [4]. Subsequently, after a series of transformations (e.g., bottleneck layer), the inter-channel dependencies are captured and the global context information is combined with the local feature maps. The attention coefficient matrix is generated by a nonlinear transformation, and this matrix is multiplied with the input feature map to obtain the enhanced feature map. The model's capacity to discriminate between various classes or areas is greatly improved by this procedure, which also propagates the global context information into the high-resolution feature map while filtering out irrelevant features and preserving critical activation zones. Table 1 shows the performance comparison of the above models on different datasets

**Table 1.** Comparison of some deep learning defection network test results

| mould | data set | mAP (%) | mIoU (%) | AP50 (%) |
|---|---|---|---|---|
| FPN | COCO | 33.9 | - | 56.9 |
| FPN | COCO | 35.0 | - | 58.2 |
| FDSNet | NEY-Seg | - | 78.8 | - |
| FDSNet | Magnetic-tile-defect-datasets | - | 63.9 | - |
| FDSNet | MSD | - | 90.2 | - |
| ETDNet | NEU-DET | 46.7 | - | 80.2 |
| ETDNet | SEDD | 23.4 | - | 48.6 |
| ETDNet | AT-FDD | 19.9 | - | 36.2 |
| PGA-Net | NEU-Seg | - | 82.15 | - |
| PGA-Net | DAGM 2007 | - | 74.78 | - |
| PGA-Net | MT_defect | - | 71.31 | - |
| PGA-Net | Road_defect | - | 79.54 | - |
| Improved Mask RCNN | Severstal | 66.75 | - | 63.43 |
| MA-YOLO | AL_Dataset | 88.1 | - | - |

## 4. New ideas for model fusion design based on commonality summarization

Both intra-class variability and inter-class similarity of steel surface defects essentially require the model to take both deep and shallow features into account during feature extraction. Multi-scale modeling, attention mechanism and discriminative training strategy are the main solution paths at present. Based on the commonalities and differences of the above models, the following new ideas to solve the problem are proposed:

First, the self-contrast module is introduced to generate two positive sample views. In order for the model to learn the common features of the same type of defects, the comparison learning network increases the similarity between views of the same defects. At the same time, it decreases the similarity between pictures of different categories so that the model can learn the differences between different categories. The processed images are then subjected to pyramid feature fusion: the images are processed through different levels of convolution to separate feature maps of different resolutions, and then the feature maps of different resolutions are adjusted to the same size through up-sampling and down-sampling, and fused through 1×1 convolution.

Considering the increase in the number of model parameters after the introduction of the contrast learning module, non-repetitive middle layer features are dynamically selected for fusion in the feature fusion stage to reduce the computational effort. Subsequently, global feature vectors are extracted by global pooling and attention weights are computed for each position to enable the model to pay more attention to important global contextual information.

Through this series of operations, the model is able to first learn the common features of similar defects and the differences of dissimilar defects in a targeted manner. Then, by combining global contextual information and multi-scale feature learning, it can effectively distinguish similar features of different categories while maximizing the recognition and generalization ability of various variations of similar defects.

## 5. Conclusion

This paper systematically reviews recent deep learning-based methods for steel surface defect detection and segmentation in recent years, focusing on the challenges of intra-class variation and inter-class similarity. By analyzing the structure and comparing the performance of representative methods such as FPN, ETDNet, PGA-Net and comparative learning models, this paper summarizes the effective strategies to improve the robustness and accuracy of the models, and proposes a new scheme that integrates the advantages of multiple methods, combining contrast network with pyramid feature fusion for feature extraction. This scheme utilizes random non-repeating regularization of low-weight resolution features to speed up the computation, and introduces a contextual attention mechanism to enhance the model's ability to recognize intra-class variability and inter-class similarity.

Despite achieving significant results, this paper has limitations. For instance, due to the lack of systematic comparative analysis under unified experimental conditions, the horizontal assessment and objective judgment of the performance of each method are limited.

Looking ahead, future research include optimizing deep feature fusion structures with lightweight networks (e.g., MobileNet, EfficientNet) to improve the detection efficiency; incorporating domain adaptive and small-sample learning techniques to

improve the model's ability to migrate to new defect types; utilizing multimodal data (e.g., thermal imaging, spectral images) for feature enhancement to break through the performance bottleneck of single image information. In summary, deep learning shows a broad prospect in steel surface defect detection, but improving model accuracy, generalization ability and industrial adaptability are still the key to future research.

## References


[1] Lei, Y. -Y., & Xie, X. (2018). Improved method of Giuffre-Menegotto-Pinto hysteretic constitutive model. *Journal of Zhejiang University (Engineering Science), 52*(10), 1926- 1934.

[2] Liang, L. M., Long, P. W., & Lu, B. H. (2024). Improvement of GBS-YOLOv7t for steel surface defect detection. *Opto-Electron Engineering, 51*(5), 240044.

[3] Chan, S., Li, S., Zhang, H., Zhou, X., Mao, J., & Hong, F. (2024). Feature optimization-guided high-precision and real-time metal surface defect detection network. *Scientific reports, 14*(1), 31941.

[4] Dong, H., Song, K., He, Y., Xu, J., Yan, Y., & Meng, Q. (2020). PGA-Net: Pyramid Feature Fusion and Global Context Attention Network for Automated Surface Defect Detection. *IEEE Transactions on Industrial Informatics, 16*(12), 7448-7458.

[5] Yu, H., Song, K., Meng, Q., Yan, Y., & Yan, Y. (2020). An End-to-end Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Features. *IEEE Transactions on Instrumentation and Measurement, 69*(4), 1493-1504.

[6] Kaggle. (2019). Severstal: steel defect detection. https: //www. kaggle. com/c/severstal-steel-defect-detection/overview

[7] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936-944.

[8] Zhou, H., Yang, R., Hu, R., Shu, C., Tang, X., & Li, X. (2023). ETDNet: Efficient Transformer-Based Detection Network for Surface Defect Detection. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-14. 2525014.

[9] Zhang, C., Yu, B., & Wang, W. (2022). Steel Surface Defect Detection Based on Improved MASK RCNN. *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, 2176-2181.

[10] Jiang, L. (2023). MA-YOLO: A Method for Detecting Surface Defects of Aluminum Profiles With Attention Guidance. *IEEE Access, 11*, 71269-71286.

[11] Zhang, J., Ding, R., Ban, M., & Guo, T. (2022). FDSNeT: An Accurate Real-Time Surface Defect Segmentation Network. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3803-3807.